AD-A042 159 ILLINOIS UNIV AT URBANA-CHAMPAIGN DEPT OF COMPUTER SCIENCE F/G 12/1
GAUSSIAN ELIMINATION AND NUMERICAL INSTABILITY.(U)
APR 77 R D SKEEL AF-AFOSR-2854-75
UNCLASSIFIED UIUCDCS-R-77-862 AFOSR-TR-77-0801 NL

1 OF 1

ADA042159

END

DATE
FILMED
8-77

UIUCDCS-R-77-862 UILU-ENG 77 1714

2

3

GAUSSIAN ELIMINATION AND NUMERICAL INSTABILITY

by

Robert D. Skeel

See 1473

April 1977

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN · URBANA, ILLINOIS

ERRATUM

Interchange

pages 56 and 57,

pages 58 and 59,

pages 60 and 61.

UIUCDCS-R-77-862

GAUSSIAN ELIMINATION AND NUMERICAL INSTABILITY*

by

Robert D. Skeel

April 1977

Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801

*Supported in part by the US Air Force under grant AFOSR-75-2854.

1. *Introduction*

The book by Forsythe and Moler [1967] gives an error analysis for Gaussian elimination drawn from the work of Wilkinson and an analysis for iterative improvement due to Moler. In this paper we do a careful backward error analysis using a different idea of what it means for a perturbation to be small, namely that each datum is subject to a small relative change.

A system of n equations

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \ 1 \le i \le n,$$

in n unknowns x_i, $1 \le i \le n$, is often written in matrix notation as

$$Ax = b.$$

The solution $\hat{x}$ computed in floating point arithmetic is not generally exactly equal to x, but it still may be acceptable if it satisfies one of two criteria:

(i) It fulfills the accuracy requirements of the problem poser. This usually involves some measure of how far $\hat{x}$ is from x. For example, in determining the coefficients of an interpolating polynomial it is some norm of the residual $r = A(x - \hat{x})$ which is of concern. More often though it is some norm of the error $\hat{x} - x$ which is of concern.

(ii) It is the exact solution of a problem which differs from the given problem by less than the uncertainty in the data, so that it is conceivable that $\hat{x}$ exactly solves the original problem. Uncertainty in the data is generally present if only because of the roundoff error introduced when the numbers are put into the computer.

It is this second criterion which motivates backward error analysis.

In §2 we give mathematical form to the informal discussion of ill conditioning found in Hamming's [1971] book *Introduction to Applied Numerical Analysis.* The condition of a system is the sensitivity of the solution to uncertainty in the data, and this sensitivity is often measured by a condition number. The usual definition of the condition number is based on the idea that the uncertainties in the data are roughly the same size in an absolute sense. However, if we suppose that these uncertainties have roughly the same relative size, then we obtain

$$\frac{\| |A^{-1}| |A| |x| + |A^{-1}| |b| \|}{\|x\|}$$

as the condition number of a linear system where $\|\circ\|$ is the max norm. There are at least a couple of reasons for believing that this second approach is more realistic. First, most numerical computations are done in floating point rather than fixed point arithmetic, and for floating point computation the conversion of data to machine represented to numbers results in errors of the same relative size. Second, measurement errors are usually more nearly the same in relative size than in absolute size.

In §3 we pose the question: What is the least amount by which A and b must be perturbed so that $\hat{x}$ exactly solves the perturbed problem? The answer to this question turns out to be

$$\max_i \frac{|(b - A\hat{x})_i|}{(|b| + |A||\hat{x}|)_i} ,$$

which we call the "backward error." If this is less than the unit roundoff error u, then the second acceptability criterion is satisfied. And if it can be shown that an algorithm produces a backward error that is always bounded by some fixed multiple K(n)u of the unit roundoff error, then by increasing the precision of the intermediate results by a factor K(n), the second acceptability criterion can be met. An algorithm with this desirable property is said to be stable. If an algorithm is only stable for infinitesimal values of u, it is said to be asymptotically stable.

The stability of Gaussian elimination with row pivoting (usually called partial pivoting) is examined in §4. By using an example of Hamming's it is shown that row pivoting is not asymptotically stable even for systems with equilibrated matrices. Then by means of a careful error analysis performed in Appendix A a bound on the backward error is obtained which contains the quantity

$$\frac{\max_i (|D_1^{-1}A||\hat{x}|)_i}{\min_j (|D_1^{-1}A||\hat{x}|)_j}$$

where D_1^{-1} is the matrix of row scaling factors. This quantity is minimized by choosing

$$D_1 = \mathrm{diag}(|A||\hat{x}|),$$

which calls for the i-th row to be divided by $|a_{i1}\hat{x}_1| + |a_{i2}\hat{x}_2| +\ldots+ |a_{in}\hat{x}_n|$. It is shown that with such a choice for D_1 row pivoting would be stable. Of course this is impractical, which explains Stewart's [1973, p. 158] observation that "In spite of intensive theoretical investigation, there is no satisfactory algorithm for scaling a general matrix." Nonetheless, the ratio

$$\frac{\max_i (|A||\hat{x}|)_i}{\min_j (|A||\hat{x}|)_j}$$

is an excellent *a posteriori* measure of how poorly scaled the system is.

Sometimes, programming considerations (Sherman [1976]) call for the use of column pivoting instead of row pivoting, where by column pivoting we mean that columns are interchanged so that each pivot is the largest in its row. In §5 it is shown that column pivoting could be made stable if it were somehow possible to scale the columns with the matrix of scale factors

$$D_2 = \mathrm{diag}(|\hat{x}|).$$

This calls for each column to be multiplied by its corresponding computed solution value. A measure of ill scaling is given by

$$\max_i \frac{(|A|e)_i \, \|\hat{x}\|}{(|A||\hat{x}|)_i}$$

where e is the vector of all ones.

Row pivoting may be regarded as the generalization of complete pivoting in which the ordering of the columns is arbitrary, and similarly column pivoting as the generalization in which the ordering of the rows is arbitrary. From this observation it follows that the results of both §4 and §5 apply to complete pivoting. However, one suspects that the error of complete pivoting satisfies an error bound which is appreciably better than simply the smaller of the bounds for row and column pivoting.

In §6 iterative improvement is examined in the hope that the poor stability properties of Gaussian elimination can be corrected. It is shown that a single iteration of iterative improvement performed in single precision is enough to make Gaussian elimination asymptotically stable.

Before proceeding, it might be interesting to demonstrate the instability of complete pivoting with a simple 2 × 2 system of equations. Consider Ax = b where

$$A = \begin{bmatrix} 3 & 3 \\ -1 & 0 \end{bmatrix} \qquad \text{and } b = \begin{bmatrix} 1 \\ 0 \end{bmatrix} .$$

The coefficient matrix A is equilibrated according to the definition of Forsythe and Moler [1967, p. 45]. Using rounded t digit decimal (floating point) arithmetic the elimination step yields A'x = b' where

$$A' = \begin{bmatrix} 3 & 3 \\ 0 & .99\cdots 9 \end{bmatrix} \quad \text{and } b' = \begin{bmatrix} 1 \\ .33\cdots 3 \end{bmatrix} ,$$

and so the computed solution

$$\hat{x} = \begin{bmatrix} .33\cdots 3 \times 10^{-t} \\ .33\cdots 3 \end{bmatrix} .$$

The backward error is determined by considering perturbed problems of the form

$$\begin{bmatrix} 3(1 + \delta_{11}) & 3(1 + \delta_{12}) \\ -(1 + \delta_{21}) & 0 \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} = \begin{bmatrix} 1 + \delta_{13} \\ 0 \end{bmatrix}$$

and choosing the relative changes δ_{ij} *so as the minimize the maximum* $|\delta_{ij}|$. In this case, δ_{21} must be chosen to be -1, and so the backward error is 100% regardless of the precision t.

2. *Condition of Linear Systems*

The condition of a problem is the sensitivity of its solution to uncertainties in the problem data. The importance of this concept is that it indicates the amount of accuracy that one should reasonably expect for the solution of a problem with inexact data. And even for problems with exact data, the conversion of the numbers to the computer's floating point number base usually introduces errors.

As the measure of the condition of a problem we take the maximum amount by which an infinitesimal perturbation in the problem data can be amplified in the solution. More precisely if ξ denotes the given problem data and $\phi(\tilde{\xi})$ denotes the solution of a problem with data $\tilde{\xi}$, then we define the *condition number* to be

$$\overline{\lim_{\tilde{\xi}\to\xi}} \ \frac{\text{relative distance from } \phi(\tilde{\xi}) \text{ to } \phi(\xi)}{\text{relative distance from } \xi \text{ to } \xi} \ . \qquad (2.1)$$

In the case where ξ and $\phi(\xi)$ are scalars, the condition number is the absolute value of the relative derivative, namely

$$\left|\frac{\xi\phi'(\xi)}{\phi(\xi)}\right|$$

(cf. Bauer [1974]). For linear algebraic systems $Ax = b$ we have $\xi = (A, b)$ and $\phi(\xi) = A^{-1} b$. (In roundoff analysis the number of equations n is not considered to be part of the problem data; rather we take the point of view that each value of n defines a separate class of problems.) There are two crucial matters that have to be settled: (i) how to define relative distance in the problem space, and (ii) how to define relative distance in the solution space.

Any problem which is to be solved in an approximate sense is incomplete unless there is also some "metric" specified for measuring how good the approximation is. It is this metric that should be used in defining the "relative distance from $\phi(\tilde{\xi})$ to $\phi(\xi)$." Often this metric measures how close the approximate solution is to the true solution; in other cases it measures how well the approximate solution satisfies the problem. In this section we choose to consider the problem of approximating A^{-1} b rather than that of approximately solving $Ax = b$. The ratio

$$\frac{\|\tilde{x} - x\|}{\|x\|}$$

where $\| \bullet \|$ denotes the max norm is an adequate measure of relative distance for most purposes if the unknowns are appropriately scaled.

The question of how to measure the "relative distance from $\tilde{\xi}$ to ξ" is more difficult to answer because a completely specified approximation problem need not include a metric for the problem space. However, there is one metric which is always safe to use, namely the componentwise relative error

$$\max_i \frac{|\tilde{\xi}_i - \xi_i|}{|\xi_i|} .$$

If the value of this quantity is small, then $\tilde{\xi}$ is close to ξ by any reasonable standard, especially in view of the fact that putting data into the computer results in small componentwise errors. This metric has another advantage in that it is always meaningful regardless of the physical dimensions of the problem data and thus it is independent of possibly arbitrary choice of units for the data. For these reasons we take as our measure of relative distance the smallest $\varepsilon \geq 0$ such that

$$|\tilde{a}_{ij} - a_{ij}| \leq \varepsilon |a_{ij}| \text{ and } |\tilde{b}_i - b_i| \leq \varepsilon |b_i| .$$

This seems to be consistent with the ideas expressed in Hamming's [1971] book *Introduction to Applied Numerical Analysis*. On page 117 it is stated that

> "The term 'ill-conditioned' is ill defined. The vague idea is that small changes in the initial system can produce large changes in the final result. If we are to take floating point seriously, then we should say 'relatively small changes' and 'relatively large changes'."

and on page 122 it is stated that

> "...the system is indeed ill conditioned because, no matter how we try, we are unable to solve the system so that the answer is not sensitive to small changes in the original coefficients."

Thus it seems that by "relatively small changes in the initial system" Hamming means "relatively small changes in the *coefficients* of the initial system." A similar thought is expressed by Kahan [1966, p. 795]. It is worth mentioning that this approach has the advantage of forcing the perturbed matrix $\tilde{A}$ to have the same sparsity structure as the original matrix A, making it more plausible to regard $\tilde{A}$ as the result of perturbing the original physical problem.

Having chosen our metrics, we are in a position to determine the condition number of a system $Ax = b$. We begin by obtaining bounds on the uncertainty in the solution due to the uncertainty in A and b. Bounds of this type also appear in Bauer [1966]. Our notation uses inequalities between arrays to mean inequality of the corresponding components. The absolute value of an array is also to be understood in a componentwise sense.

THEOREM 2.1. *Let* $Ax = b$ *and* $(A + \delta A)(x + \delta x) = b + \delta b$ *where* $|\delta A| \le \varepsilon|A|$ *and* $|\delta b| \le \varepsilon|b|$. *Then*

$$\frac{\|\delta x\|}{\|x\|} \le \varepsilon \frac{\|\,|A^{-1}||A||x| + |A^{-1}||b|\,\|}{(1 - \varepsilon\|\,|A^{-1}||A|\,\|)\|x\|}$$

provided that the denominator is positive.

PROOF. We have that

$$\delta x = A^{-1}\delta A(x + \delta x) + A^{-1}\delta b \;, \tag{2.2}$$

and so

$$|\delta x| \leq |A^{-1}||\delta A|(|x| + |\delta x|) + |A^{-1}||\delta b|$$
$$\leq \varepsilon\ |A^{-1}||A|(|x| + |\delta x|) + \varepsilon\ |A^{-1}||b|\ .$$

Therefore

$$\|\delta x\| \leq \varepsilon\ \||A^{-1}||A||x| + |A^{-1}||b|\| + \varepsilon\||A^{-1}||A|\|\ \|\delta x\|\ . \qquad \text{Q.E.D.}$$

Note. Bauer [1966] shows that the bound of Theorem 2.1 can be improved by replacing $(1 - \varepsilon\ \||A^{-1}||A|\|)^{-1}$ by $\|(I - \varepsilon\ |A^{-1}||A|)^{-1}\|$, and so it is not necessary that $\varepsilon\ \||A^{-1}||A|\| < 1$ but only that the spectral radius of $\varepsilon\ |A^{-1}||A|$ be less than 1.

THEOREM 2.2. *Let* $Ax = b$. *Then there exist* δA *and* δb *such that* $|\delta A| = \varepsilon\ |A|$, $|\delta b| = \varepsilon\ |b|$, *and the solution* $x + \delta x$ *of* $(A + \delta A)(x + \delta x) = b + \delta b$ *satisfies*

$$\frac{\|\delta x\|}{\|x\|} \geq \frac{\||A^{-1}||A||x| + |A^{-1}||b|\|}{(1 + \varepsilon\ \||A^{-1}||A|\|)\|x\|}\ .$$

PROOF. Let ℓ be such that

$$(|A^{-1}||A||x| + |A^{-1}||b|)_\ell = \||A^{-1}||A||x| + |A^{-1}||b|\|\ .$$

Define δA and δb by

$$\delta a_{jk} = \mathrm{sgn}(\alpha_{\ell j} x_k)\ \varepsilon\ |a_{jk}|$$

and

$$\delta b_j = \mathrm{sgn}(\alpha_{\ell j})\ \varepsilon\ |b_j|$$

where $A^{-1} = (\alpha_{ij})$. Then

$$(A^{-1}\delta Ax + A^{-1}\delta b)_\ell = \sum_j \sum_k \alpha_{\ell j}\ \delta a_{jk}\ x_k + \sum_j \alpha_{\ell j}\ \delta b_j$$
$$= \varepsilon\ (|A^{-1}||A||x| + |A^{-1}||b|)_\ell$$
$$= \varepsilon\ \||A^{-1}||A||x| + |A^{-1}||b|\|\ ,$$

but from (2.2)

$$(A^{-1}\delta Ax + A^{-1}\delta b)_{\ell} = (\delta x - A^{-1}\delta A\delta x)_{\ell} \ ,$$

and so

$$\varepsilon \, \| |A^{-1}||A||x| + |A^{-1}||b| \| \leq \|\delta x\| + \varepsilon \, \| |A^{-1}||A| \| \, \|\delta x\| \ . \qquad \text{Q.E.D.}$$

THEOREM 2.3. *The condition number, as defined by* (2.1), *of a linear algebraic system* Ax = b *is*

$$\frac{\| |A^{-1}||A||x| + |A^{-1}||b| \|}{\|x\|} \ .$$

PROOF. The condition number of a linear algebraic system Ax = b is

$$\overline{\lim_{\varepsilon(\delta A,\delta b) \to 0}} \ \frac{\|\delta x\|/\|x\|}{\varepsilon(\delta A,\delta b)}$$

where $\varepsilon(\delta A,\delta b) = \min\{\varepsilon \geq 0\colon |\delta A| \leq \varepsilon|A|,\ |\delta b| \leq \varepsilon|b|\}$ and δx satisfies $(A + \delta A)(x + \delta x) = b + \delta b$. Consider any sequence $(\delta A_m, \delta b_m)$ for which $\varepsilon(\delta A_m, \delta b_m) \to 0$ as $m \to \infty$. By Theorem 2.1 we have

$$\frac{\|\delta x\|}{\|x\|} \leq \varepsilon(\delta A_m,\delta b_m) \frac{\| |A^{-1}||A||x| + |A^{-1}||b| \|}{(1 - \varepsilon(\delta A_m,\delta b_m) \, \| |A^{-1}||A| \|) \, \|x\|}$$

for sufficiently large m. Therefore

$$\overline{\lim_{m \to \infty}} \ \frac{\|\delta x_m\|/\|x\|}{\varepsilon(\delta A_m,\delta b_m)} \leq \frac{\| |A^{-1}||A||x| + |A^{-1}||b| \|}{\|x\|} \ ,$$

which gives an upper bound on the condition number. Let ε_m be a sequence converging to zero. By Theorem 2.2 there exists a sequence $(\delta A_m, \delta b_m)$ such that $\varepsilon(\delta A_m, \delta b_m) = \varepsilon_m$ and

$$\frac{\|\delta x_m\|}{\|x\|} \geq \varepsilon_m \frac{\| |A^{-1}||A||x| + |A^{-1}||b| \|}{(1 + \varepsilon_m \| |A^{-1}||A| \|) \, \|x\|} \ .$$

Therefore

$$\overline{\lim_{m \to \infty}} \ \frac{\|\delta x_m\|/\|x\|}{\varepsilon_m} \geq \frac{\| |A^{-1}||A||x| + |A^{-1}||b| \|}{\|x\|} \ ,$$

which gives a lower bound on the condition number. Q.E.D.

In subsequent sections of this paper, we will consider the effects of perturbing only the elements of the coefficient matrix.

THEOREM 2.4. *Let* $Ax = b$ *and* $(A + \delta A)(x + \delta x) = b$ *where* $|\delta A| \le \varepsilon |A|$. *Then*

$$\frac{\|\delta x\|}{\|x\|} \le \frac{\varepsilon \| |A^{-1}| |A| |x| \|}{(1 - \varepsilon \| |A^{-1}| |A| \|) \|x\|} .$$

PROOF. Similar to that of Theorem 2.1. Q.E.D.

THEOREM 2.5. *Let* $Ax = b$. *Then there exists* δA *such that* $|\delta A| = \varepsilon |A|$ *and such that the solution* $x + \delta x$ *of* $(A + \delta A)(x + \delta x) = b$ *satisfies*

$$\frac{\|\delta x\|}{\|x\|} \ge \frac{\varepsilon \| |A^{-1}| |A| |x| \|}{(1 + \varepsilon \| |A^{-1}| |A| \|) \|x\|} .$$

PROOF. Similar to that of Theorem 2.2. Q.E.D.

It follows from these last two theorems that when only A is subject to uncertainty the condition number is

$$\mathrm{Cond}(A, x) = \frac{\| |A^{-1}| |A| |x| \|}{\|x\|} .$$

Since $\| |A^{-1}| |A| |x| \| \le \| |A^{-1}| |A| |x| + |A^{-1}| |b| \| \le 2\| |A^{-1}| |A| |x| \|$, Cond(A, x) is also adequate for the case where both A and b are subject to uncertainty. A similar quantity

$$\kappa(A, x) = \frac{\|A^{-1}\| \sum_j \|Ae_{(j)}\| \, |x_j|}{\|x\|} .$$

is used by Van der Sluis [1970a], which he calls [1970b] the "condition number of the solution." Here $e_{(j)}$ denotes the j-th unit vector.

The condition number of a matrix A could be defined as the maximum value of Cond(A, x), which is achieved with $x = e = (1, 1, \ldots, 1)^T$. Thus

$$\mathrm{Cond}(A) = \mathrm{Cond}(A, e) = \| |A^{-1}| |A| \|.$$

This quantity is more satisfying as a measure of ill condition than the usual $\mathrm{cond}(A) = \|A^{-1}\| \|A\|$ for a couple of reasons. First, the matrix

$|A^{-1}||A|$ is a mapping of the solution space into itself, which means that the quantity $\| |A^{-1}||A| \|$ can be defined entirely in terms of the solution space norm. Whereas $\text{cond}(A) = \|A^{-1}\| \, \|A\|$ is defined in terms of a solution space norm and a residual space norm, which seems quite unnecessary. Second, the quantity Cond(A) is invariant under row scaling. Multiplying a system of equations by a diagonal matrix does not change the problem in any fundamental way. For example, all systems $Dx = b$ where D is diagonal are well conditioned. Accordingly, we have that $\text{Cond}(D) = 1$; whereas $\text{cond}(D)$ can be arbitrarily large.

Example. According to Hamming [1971, p. 120], the system $Ax = b$ is well conditioned where

$$A = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 2\varepsilon & 2\varepsilon \\ 1 & 2\varepsilon & -\varepsilon \end{bmatrix}, \qquad b = \begin{bmatrix} 3 + 3\varepsilon \\ 6\varepsilon \\ 2\varepsilon \end{bmatrix}.$$

The inverse of the coefficient matrix and the solution are given below:

$$A^{-1} = \frac{1}{1-1.8\varepsilon} \begin{bmatrix} -.6\varepsilon & .4 & .2 \\ .4 & -.1\varepsilon^{-1}-.3 & .2\varepsilon^{-1}-.6 \\ .2 & .2\varepsilon^{-1}-.6 & -.4\varepsilon^{-1}-.6 \end{bmatrix}, \qquad x = \begin{bmatrix} \varepsilon \\ 1 \\ 1 \end{bmatrix}.$$

Hence

$$|A^{-1}||A| = \frac{1}{1-1.8\varepsilon} \begin{bmatrix} 1+1.8\varepsilon & 2.4\varepsilon & 1.6\varepsilon \\ .4\varepsilon^{-1}+1.2 & 1.4-.6\varepsilon & .8 \\ .8\varepsilon^{-1} & 1.6 & 1-.6\varepsilon \end{bmatrix},$$

and

$$|A^{-1}||A||x| + |A^{-1}||b| = \frac{1}{1-1.8\varepsilon}\begin{bmatrix} 9.6\varepsilon + 3.6\varepsilon^2 \\ 4.8 + 2.4\varepsilon \\ 6 - 2.4\varepsilon \end{bmatrix},$$

which shows that the system is well conditioned. However,

$$\text{Cond}(A) = \frac{.8\varepsilon^{-1} + 2.6 - .6\varepsilon}{1-1.8\varepsilon},$$

which indicates that the system would be ill conditioned for some different right hand side b, and in fact, Hamming [1971, p. 122] gives such an example.

3. *Stability of Algorithms for Linear Systems*

Let $\hat{+}$, $\hat{-}$, $\hat{\times}$, $\hat{/}$ denote the floating point operations corresponding to +, −, ×, /. Every reference to a floating point result $x \mathbin{\hat{\circ}} y$ carries with it the assumption that x, $\hat{\circ}$, and y are such that the result is well defined. Nothing is assumed about the floating point arithmetic except that the relative roundoff error is bounded by $u/(1 + u)$ where the unit roundoff error u is a small positive number; that is,

$$x \mathbin{\hat{\circ}} y = (x \circ y)(1 + \delta)$$

for some δ depending on x, $\hat{\circ}$, and y which satisfies

$$|\delta| \le \frac{u}{1 + u} .$$

It follows from the above condition that

$$x \mathbin{\hat{\circ}} y = \frac{x \circ y}{1 + \delta'}$$

where $|\delta'| \le u$. Note that for rounding $u = \frac{1}{2} \beta^{1-t}$ and for chopping $u = \beta^{1-t}$ where β is the base and t is the number of base β digits in the fraction of the floating point numbers.

For any computed solution $\hat{x}$ we define the *relative backward error* to be the smallest real number $\eta_{(3)}$ such that

$$(A + \delta A)(\hat{x} - \delta x) = b + \delta b$$

for some δA, δb, and δx with $|\delta A| \le \eta_{(3)}|A|$, $|\delta b| \le \eta_{(3)}|b|$, and $|\delta x| \le \eta_{(3)}|\hat{x} - \delta x|$. The backward error can be interpreted in the following way: The computed solution $\hat{x}$ is the rounded solution of a problem with rounded data where $\eta_{(3)}$ is the maximum relative roundoff error. Thus, if $\eta_{(3)}$ is no larger than the unit roundoff error u, then our solution is as good as our data deserve; otherwise, improved accuracy may be justified, perhaps by using iterative improvement. Further motivation for this definition is given in Miller [1975] and Bauer [1974]. If a solution cannot be computed because the matrix is nearly singular, then the backward error is defined to be the smallest real number $\eta_{(3)}$ such that A + δA is singular for some $|\delta A|$ with $|\delta A| \le \eta_{(3)}|A|$.

Stability of the algorithm means that there exists a *stability constant* K(n) and a *stability threshold* $\bar{u}(n) > 0$ both independent of the problem data (A, b) such that the relative backward error

$$\eta_{(3)} \le K(n)u$$

provided that $u \le \bar{u}(n)$. A weaker concept *asymptotic stability* allows the threshold $\bar{u}(n)$ to be data dependent. These two types of stability are the same as the "backward stability" and "asymptotic backward stability" used by Miller [1972].

The backward error $\eta_{(3)}$ is not easy to determine, and for this reason we introduce two variants of the backward error which are easier to compute. Let $\eta_{(2)}$ be the smallest real number such that

$$(A + \delta A)\hat{x} = b + \delta b$$

for some δA and δb with $|\delta A| \le \eta_{(2)}|A|$ and $|\delta b| \le \eta_{(2)}|b|$. Let $\eta_{(1)}$ be the smallest real number such that

$$(A + \delta A)\hat{x} = b$$

for some δA with $|\delta A| \le \eta_{(1)}|A|$. Naturally $\eta_{(3)} \le \eta_{(2)} \le \eta_{(1)}$.

THEOREM 3.1. *Let* $I = \{i: \ (|A||\hat{x}| + |b|)_i = 0\}$. *The backward error*

$$\eta_{(2)} = \begin{cases} \max\limits_{i \notin I} \dfrac{|(b - A\hat{x})_i|}{(|A||\hat{x}| + |b|)_i} & \textit{if } (b - A\hat{x})_i = 0 \textit{ for } i \in I, \\ \\ +\infty & \textit{otherwise}. \end{cases}$$

PROOF. First consider the case where $(b - A\hat{x})_i \ne 0$ for some $i \in I$. Suppose that $\eta_{(2)} < +\infty$. Then there exist δA and δb such that $(A + \delta A)\hat{x} = b + \delta b$ where $|\delta A| \le \eta_{(2)}|A|$ and $|\delta b| \le \eta_{(2)}|b|$. We have

$$\begin{aligned} |b - A\hat{x}| &= |\delta A\hat{x} - \delta b| \\ &\le |\delta A||\hat{x}| + |\delta b| \\ &\le \eta_{(2)} \; (|A||\hat{x}| + |b|), \end{aligned} \qquad (3.1)$$

which is impossible since $(b - A\hat{x})_i \neq 0$ and $(|A||\hat{x}| + |b|)_i = 0$. Therefore $\eta_{(2)} = +\infty$. Second consider the case where $(b - A\hat{x})_i = 0$ for all $i \in I$. To obtain an upper bound on $\eta_{(2)}$, consider the choice

$$\delta a_{ij} = \begin{cases} \operatorname{sgn}(\hat{x}_j)|a_{ij}| \dfrac{(b - A\hat{x})_i}{(|A||\hat{x}| + |b|)_i} & i \notin I, \\ 0 & i \in I, \end{cases}$$

and

$$\delta b_i = \begin{cases} -|b_i| \dfrac{(b - A\hat{x})_i}{(|A||\hat{x}| + |b|)_i} & i \notin I, \\ 0 & i \in I. \end{cases}$$

We have $\delta A\hat{x} - \delta b = b - A\hat{x}$ or $(A + \delta A)\hat{x} = b + \delta b$, and so

$$\eta_{(2)} \le \max_{i \notin I} \frac{|(b - A\hat{x})_i|}{(|A||\hat{x}| + |b|)_i} .$$

Since $\eta_{(2)} < +\infty$, equation (3.1) must hold; and so

$$\eta_{(2)} \ge \max_{i \notin I} \frac{|(b - A\hat{x})_i|}{(|A||\hat{x}| + |b|)_i} . \qquad \text{Q.E.D.}$$

THEOREM 3.2. *Let* $I = \{i : (|A||\hat{x}|)_i = 0\}$. *The backward error*

$$\eta_{(1)} = \begin{cases} \max\limits_{i \notin I} \dfrac{|(b - A\hat{x})_i|}{(|A||\hat{x}|)_i} & \textit{if} \;\; (b - A\hat{x})_i = 0 \;\; \textit{for} \;\; i \in I, \\ +\infty & \textit{otherwise}. \end{cases}$$

PROOF. Similar to Theorem 3.1. Q.E.D.

Remark. Similar types of results apply to other problems; for example, the backward error for a polynomial equation $a_0 x^n + a_1 x^{n-1} + \ldots + a_n = 0$ is given by

$$\eta_{(2)} = \frac{|a_0 \hat{x}^n + a_1 \hat{x}^{n-1} + \ldots + a_n|}{|a_0 \hat{x}^n| + |a_1 \hat{x}^{n-1}| + \ldots + |a_n|} .$$

The following theorem gives bounds on the relative backward error $\eta_{(3)}$ in terms of the more easily computed $\eta_{(1)}$ and $\eta_{(2)}$. These bounds show that the η's are roughly the same size when the backward error is small, and so in the remainder of the paper only the quantity $\eta_{(1)}$ is used, which we denote simply by η.

THEOREM 3.3. *The three types of backward error satisfy*

$$\frac{\eta_{(2)}}{2 + \eta_{(2)}} \leq \eta_{(3)} \leq \eta_{(2)} , \tag{3.2}$$

$$\frac{\eta_{(1)}}{3 + \frac{5}{3}\eta_{(1)}} \leq \eta_{(3)} \leq \eta_{(1)} , \tag{3.3}$$

$$\frac{\eta_{(1)}}{2 + \eta_{(1)}} \leq \eta_{(2)} \leq \eta_{(1)} . \tag{3.4}$$

PROOF. The second inequalities of (3.2), (3.3), and (3.4) are obvious. The first inequality of (3.2) is obvious if $\eta_{(3)} \geq 1$. Hence assume $\eta_{(3)} < 1$. There exist δA, δb, δx such that

$$(A + \delta A)(\hat{x} - \delta x) = b + \delta b$$

where $|\delta A| \leq \eta_{(3)}|A|$, $|\delta b| \leq \eta_{(3)}|b|$, $|\delta x| \leq \eta_{(3)}|\hat{x} - \delta x|$. Hence

$$\begin{aligned} |b - A\hat{x}| &= |\delta A(\hat{x} - \delta x) - A\delta x - \delta b| \\ &\leq 2\eta_{(3)}|A||\hat{x} - \delta x| + \eta_{(3)}|b| . \end{aligned} \tag{3.5}$$

It is easily shown

$$|\hat{x} - \delta x| \leq \frac{1}{1 - \eta_{(3)}} |\hat{x}|, \tag{3.6}$$

and so

$$|b - A\hat{x}| \leq \frac{2\eta_{(3)}}{1 - \eta_{(3)}} (|A||\hat{x}| + |b|).$$

Therefore $\eta_{(2)} \leq 2\eta_{(2)}/(1 - \eta_{(3)})$ which verifies (3.2). The first inequality of (3.3) is obvious if $\eta_{(3)} \geq \frac{3}{5}$. Hence assume $\eta_{(3)} < \frac{3}{5}$. From (3.5)

$$|b - A\hat{x}| \leq 2\eta_{(3)}|A||\hat{x} - \delta x| + \eta_{(3)}|b - A\hat{x}| + \eta_{(3)}|A||\hat{x}|,$$

and using (3.6) gives

$$(1 - \eta_{(3)})|b - A\hat{x}| \leq (\frac{2\eta_{(3)}}{1 - \eta_{(3)}} + \eta_{(3)})|A||\hat{x}|.$$

Therefore

$$\eta_{(1)} \leq \frac{\eta_{(3)} \; (3 - \eta_{(3)})}{(1 - \eta_{(3)})^2}$$

$$\leq \frac{3\eta_{(3)} \; (1 - \frac{1}{3}\eta_{(3)})}{(1 - \frac{5}{3}\eta_{(3)})(1 - \frac{1}{3}\eta_{(3)})} \; ,$$

which proves (3.3). The first unequality of (3.4) is obvious if $\eta_{(2)} \geq 1$. Hence assume $\eta_{(2)} < 1$. We have

$$|b - A\hat{x}| \leq \eta_{(2)}(|b| + |A||\hat{x}|) \leq \eta_{(2)}(|b - A\hat{x}| + 2|A||\hat{x}|),$$

and so

$$|b - A\hat{x}| \leq \frac{2\eta_{(2)}}{1 - \eta_{(2)}} \; |A||\hat{x}|.$$

Therefore $\eta_{(1)} \leq 2\eta_{(2)}/(1 - \eta_{(2)})$, which implies (3.4). Q.E.D.

A good algorithm should (i) return an acceptable answer most of the time (robustness), and (ii) signal failure whenever it does not return an acceptable answer (reliability). We formally define an algorithm to be reliable if there exist $K(n)$ and $\bar{u}(n)$ such that for any (A, b) and any $u \leq \bar{u}(n)$ either the algorithm computes an answer with $\eta \leq K(n)u$ or the algorithm signals failure.

Any algorithm for solving linear systems can be made reliable by computing the backward error with floating point arithmetic and then accepting the answer only if the computed backward error is less than a prescribed multiple of the unit roundoff error. For example, the next theorem shows that if the computed backward error $\hat{\eta} \leq Ku$, then we can conclude that $\eta \leq (K + n)ue^{(n+2)u}$.

The residual is to be computed in single precision

$$\hat{r}_i = b_i \,\hat{-}\, (\ldots(a_{i1} \,\hat{\times}\, \hat{x}_1 \,\hat{+}\, a_{i2} \,\hat{\times}\, \hat{x}_2)\ldots \,\hat{+}\, a_{in} \,\hat{\times}\, \hat{x}_n)$$

or in double precision

$$\hat{r}_i = \mathrm{fl}(b_i \,\dot{-}\, (\ldots(a_{i1} \,\dot{\times}\, \hat{x}_i \,\dot{+}\, a_{i2} \,\dot{\times}\, \hat{x}_2)\ldots \,\dot{+}\, a_{in} \,\dot{\times}\, \hat{x}_n)).$$

Here $\dot{\circ}$ denotes the double precision counterpart of $\hat{\circ}$ where it is assumed that

$$x \,\dot{\circ}\, y = (x \circ y)(1 + \delta)$$

with $|\delta| \leq u^2/(1 + u^2)$. In practice the double precision unit roundoff error is either this small (rounding in base two) or smaller. By fl(°) we mean the conversion of a double precision value to a single precision value. It is assumed that $\mathrm{fl}(x\dot{\circ}y) = (x\circ y)(1+\delta)$ with $-u/(1+u) \leq \delta \leq u$, which is true for rounding and chopping. The computed backward error $\hat{\eta}$ is determined by

$$\hat{\eta} = \max_i (|\hat{r}_i| \,\hat{/}\, (\ldots(|a_{i1} \,\hat{\times}\, \hat{x}_1| \,\hat{+}\, |a_{i2} \,\hat{\times}\, \hat{x}_2|)\ldots \,\hat{+}\, |a_{in} \,\hat{\times}\, \hat{x}_n|).$$

THEOREM 3.4. *If* $\hat{\eta}$ *is the computed value of* η, *then*

$$e^{-(n+2)u}\,\hat{\eta} - n\bar{u}\,e^{n\bar{u}} \leq \eta \leq e^{(n+2)u}\,\hat{\eta} + n\bar{u}\,e^{n\bar{u}}$$

where

$$\bar{u} = \begin{cases} u & \text{\textit{for single precision residual accumulation,}} \\ u^2 & \text{\textit{for double precision residual accumulation.}} \end{cases}$$

PROOF. Let $\hat{q}$ be the computed value of $A\hat{x}$. By the usual type of error analysis

$$\begin{aligned} |\hat{q} - A\hat{x}| &\leq [(1 + \bar{u})^n - 1]|A||\hat{x}| \\ &\leq n\bar{u}\,e^{n\bar{u}}|A||\hat{x}| \,. \end{aligned} \tag{3.7}$$

We have that

$$\hat{\eta} \geq \frac{(1 - \frac{u}{1+u})^2}{(1 + \frac{u}{1+u})^n} \max \frac{|b - \hat{q}|}{|A||\hat{x}|}$$

and

$$\hat{\eta} \le \frac{(1 + u)(1 + \frac{u}{1 + u})}{(1 - \frac{u}{1 + u})^n} \max \frac{|b - \hat{q}|}{|A||\hat{x}|}$$

where division of two vectors is defined componentwise. This reduces to

$$\frac{1}{(1 + 2u)(1 + u)^n} \hat{\eta} \le \max \frac{|b - \hat{q}|}{|A||\hat{x}|} \le \frac{(1 + 2u)^n}{(1 + u)^{n-2}} \hat{\eta},$$

from which it follows that

$$e^{-(n+2)u} \hat{\eta} \le \max \frac{|b - \hat{q}|}{|A||\hat{x}|} \le e^{(n+2)u} \hat{\eta}.$$

Using

$$|b - \hat{q}| - |\hat{q} - A\hat{x}| \le |b - A\hat{x}| \le |b - \hat{q}| + |\hat{q} - A\hat{x}| .$$

and (3.7) gives

$$|b - \hat{q}| - n\bar{u}\, e^{(n-2)\bar{u}}|A||\hat{x}| \le |b - A\hat{x}| \le |b - \hat{q}| + n\bar{u}\, e^{(n-2)\bar{u}}|A||\hat{x}|.$$

Dividing by $|A||\hat{x}|$ and taking the maximum yields

$$\max \frac{|b - \hat{q}|}{|A||\hat{x}|} - n\bar{u}\, e^{(n-2)\bar{u}} \le \eta \le \max \frac{|b - \hat{q}|}{|A||\hat{x}|} + n\bar{u}\, e^{(n-2)\bar{u}} . \qquad \text{Q.E.D.}$$

Before concluding this section, it should be mentioned that for some classes of problems it may be unreasonable to expect an algorithm to be stable. If the number of output values is fairly large compared to the number of input values, then it becomes very difficult for an algorithm to be stable because each output value must arise from the *same* perturbation of the input values. For example, Miller [1975] shows that the usual algorithm for inverting triangular matrices is unstable. Hence it seems better to use stability as a relative concept rather than an absolute concept. This idea is used by Miller [1976] in a paper entitled "Roundoff analysis by direct comparison of two algorithms."

4. *Gaussian Elimination with Row Pivoting and Scaling*

This section applies the ideas of the preceding section to Gaussian elimination with partial pivoting using row interchanges and implicit row scaling. The reciprocals of the scale factors are to be given as inputs $d_1, d_2, \ldots, d_n$ to the algorithm, and so the pivoting is done as if one were solving $D^{-1}Ax = D^{-1}b$ where $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$. To keep the notation simple, it is assumed that the equations are numbered according to their ordering after all row interchanges have been performed. The computations of the algorithm are as follows:

$$a_{ij}^{(1)} = a_{ij},$$

$$\text{for } k = 1(1)n - 1, \begin{cases} m_{ik} = a_{ik}^{(k)} \hat{/} a_{kk}^{(k)}, \quad i \geq k + 1, & (4.1) \\ a_{ij}^{(k+1)} = a_{ij}^{(k)} \hat{-} m_{ik} \hat{\times} a_{kj}^{(k)}, \quad i,j \geq k + 1, & (4.2) \end{cases}$$

$$\ell_{ij} = \begin{cases} 0 & \text{if } i < j, \\ 1 & \text{if } i = j, \\ m_{ij} & \text{if } i > j, \end{cases}$$

$$u_{ij} = \begin{cases} a_{ij}^{(i)} & \text{if } i \leq j, \\ 0 & \text{if } i > j, \end{cases}$$

$$y_1 = b_1.$$

$$\text{for } i = 2(1)n,\ y_i = b_i \hat{-} (\ldots(\ell_{i1} \hat{\times} y_1 \hat{+} \ell_{12} \hat{\times} y_2)\ldots \hat{+} \ell_{i,i-1} \hat{\times} y_{i-1}), \qquad (4.3)$$

$$\hat{x}_n = y_n \hat{/} u_{nn},$$

$$\text{for } i = n - 1(-1)1, \qquad (4.4)$$

$$\hat{x}_i = (y_i \hat{-} (\ldots(u_{i,i+1} \hat{\times} \hat{x}_{i+1} \hat{+} u_{i,i+2} \hat{\times} \hat{x}_{i+2})\ldots \hat{+} u_{in} \hat{\times} \hat{x}_n))\hat{/}u_{ii}.$$

It is assumed that the selection of a pivot is done exactly so that

$$|d_i^{-1}\ a_{ik}^{(k)}| \leq |d_k^{-1}\ a_{kk}^{(k)}|, \quad i \geq k + 1. \qquad (4.5)$$

In cases where there are more than one suitable pivot, the one with the lowest row index is chosen. The assumption of exact pivot choice avoids some minor technical difficulties, and it also makes for a sharp error bound in the case where there is no scaling.

It is important to appreciate the nature of the functional relationship between $\hat{x}$ and D. The computed solution $\hat{x}$ is a function $\xi(P)$ of the row permutation P, which in turn is a function $\Pi(D)$ of the scaling matrix D. (Note that $\Pi(D)$ is also defined for values of D which are not floating point numbers because the algorithm does not perform floating point arithmetic on D.) If $\hat{x}$ is viewed as a function defined in $(d_1, d_2, \ldots, d_n)$-space, it would be constant over regions bounded by hyperplanes passing through the origin. For example, let $\tilde{a}_{ij}^{(k)}$ be the values corresponding to a certain choice $d_i = \tilde{d}_i$ of scale factors. Then $\hat{x}$ is constant for all values of $(d_1, d_2, \ldots, d_n)$ which satisfy

$$|d_i| > \left|\frac{\tilde{a}_{ik}^{(k)}}{\tilde{a}_{kk}^{(k)}}\right| |d_k| \ , \quad i > k.$$

It is necessary for the completeness of the theory that the Gaussian elimination algorithm be extended to permit the use of zero as a scaling factor. For any diagonal matrix $D = \operatorname{diag}(d_1, d_2, \ldots, d_n)$ we let D_ε denote the matrix $\operatorname{diag}(\{d_1\}_\varepsilon, \{d_2\}_\varepsilon, \ldots, \{d_n\}_\varepsilon)$ where

$$\{d\}_\varepsilon = \begin{cases} d & \text{if } d \neq 0 \\ \varepsilon & \text{if } d = 0. \end{cases}$$

That is, D_ε is the matrix D with all the zero diagonal elements replaced by ε. The condition (4.5) is replaced by the following:

$$\lim_{\varepsilon \to 0} \left|\frac{\{d_i\}_\varepsilon^{-1} \ a_{ik}^{(k)}}{\{d_k\}_\varepsilon^{-1} \ a_{kk}^{(k)}}\right| \leq 1.$$

Let $\bar{\varepsilon} = \min\{|d_i \ a_{kk}^{(k)}/a_{ik}^{(k)}|: \ d_k = 0, \ d_i \neq 0\}$. Then it can be easily shown that for any ε with $0 < |\varepsilon| < \bar{\varepsilon}$ the scale matrix D_ε has the same effect on the choice of pivots as does D.

Unfortunately, Gaussian elimination with pivoting is unstable. This instability arises in the decomposition stage when the quantities

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} \mathbin{\hat{-}} m_{ik} \mathbin{\hat{\times}} a_{kj}^{(k)}$$

are being formed. If $|m_{ik}\, a_{kj}^{(k)}| >> |a_{ij}^{(k)}|$, then the error in $a_{ij}^{(k+1)}$ is not very small relative to $a_{ij}^{(k)}$, and so in our backward error analysis we cannot throw this error back into $a_{ij}^{(k)}$. The extreme case occurs when $a_{ij}^{(k)} = 0$ and $a_{ij}^{(k+1)} \neq 0$, which is commonly called "fill in." For sparse systems of equations it is quite common to order the rows so as to avoid fill in. This reduces computational cost, and it apparently may also contribute to stability.

The instability of Gaussian elimination has been pointed out by Hamming, who on page 119 of his book [1971] announces the

> "Theorem Pivoting can take a well-conditioned system into an ill-conditioned system of simultaneous linear equations."

and on page 123 states

> "We have not justified the pivoting method; rather we have shown that it is an 'old wives' tale.' But like most old wives' tales, it is a mixture of truth and mystic faith."

To prove his theorem, Hamming uses the example discussed at the end of §2. For this example, one elimination step with partial or complete pivoting yields the system A'x = b' where

$$A' = \begin{bmatrix} 3 & 2 & 1 \\ 0 & \frac{-4}{3} + 2\varepsilon & \frac{-2}{3} + 2\varepsilon \\ 0 & \frac{-2}{3} + 2\varepsilon & \frac{-1}{3} - \varepsilon \end{bmatrix}, \quad b' = \begin{bmatrix} 3 + 3\varepsilon \\ -2 + 4\varepsilon \\ -1 + \varepsilon \end{bmatrix} \qquad (4.6)$$

assuming exact arithmetic. This problem is ill conditioned for small ε, since

$$\mathrm{Cond}(A',x) = \frac{.8\varepsilon^{-1} - 3 + 3\varepsilon}{1 - 1.8\varepsilon} .$$

If the elimination were performed in floating point arithmetic, then a slight perturbation of (4.6) could result, which may have a solution which differs from the true solution by an amount proportional to ε^{-1}. This kind of error could not arise from slightly perturbing the original problem because it has a condition number of about 6. For example, suppose that the computed righthand side of (4.6) was

$$\hat{b}' = \begin{bmatrix} 3 + 3\varepsilon \\ -2 + 4\varepsilon - \frac{u}{4} \\ -1 + \varepsilon + \frac{u}{2} \end{bmatrix}$$

and everything else was exact. Then

$$\hat{x} = \begin{bmatrix} \varepsilon \\ 1 - \varepsilon^{-1}\ u/8 \\ 1 + \varepsilon^{-1}\ u/4 \end{bmatrix},$$

and by Theorem 3.2 the backward error is $u/(8\varepsilon + u)$.

A related observation was made by Gear [1975]:

> "It might be possible to say that δA represents a perturbation to the original physical problem if the sparsity structure of δA were the same as that of A. Unfortunately, we will show that such a demand on the structure of δA can lead to very large bounds on $||\delta A||$, bounds probably dependent on the condition number of A."

This was supported by the example

$$A = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 0 & \varepsilon & 0 & 0 \\ 0 & 0 & \varepsilon & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 2 \end{bmatrix}, \quad x = \begin{bmatrix} 1 \\ \varepsilon^{-1} \\ \varepsilon^{-1} \\ 1 \end{bmatrix},$$

for which Cond(A) = 4.

In one of the theorems that follow, it is shown that the situation is not quite as bad as Gear suggests. The bounds on $||\delta A||$ depend not on how ill conditioned the problem is but on how badly scaled the equations are. We begin by obtaining a totally *a priori* error bound for Gaussian elimination. The proof is modeled after that of Forsythe and Moler [1967], which is mostly borrowed from Wilkinson [1963]. However, our error bound, like that of Van der Sluis [1970b], is more informative than that of Forsythe and Moler in that it distinguishes among the columns of ΔA.

THEOREM 4.1. *Let the vector* $\hat{x}$ *be computed by Gaussian elimination with row pivoting and row scaling where* $D = \text{diag}(d_1, d_2, \ldots, d_n)$ *is the matrix of reciprocal scale factors. Then there exists* ΔA *such that*

$$(A + \Delta A)\hat{x} = b$$

with

$$|| |D_\varepsilon^{-1} \Delta A|z|| \leq \chi(n)u|| |D_\varepsilon^{-1} A|z||$$

for arbitrary $z \geq 0$ *and arbitrary* ε *satisfying* $0 < |\varepsilon| < \bar{\varepsilon}$ *where*

$$\chi(n) = [19 \cdot 2^{n-2} - n - 8]e^{2nu}.$$

PROOF. See Appendix A. Q.E.D.

Note 1. The factor e^{2nu} appears in the Forsythe and Moler book as the constant 1.01. The advantage of e^{2nu} is that it indicates the nature of the higher order effects and it does not require placing some arbitrary restriction on the size of nu.

Note 2. It is actually possible to show that $|\Delta A| \leq L_B|A|$ for some lower triangular matrix L_B, although the best possible L_B is somewhat complicated.

With $z = e$ in Theorem 4.1 we get the usual type of bound

$$||D_\varepsilon^{-1} \Delta A|| \leq \chi(n)u \, ||D_\varepsilon^{-1} A|| \, .$$

The bound of this theorem is practically always extremely pessimistic. However, there are cases where this bound can be attained in the limit as $u \to 0$.

THEOREM 4.2. *There exists a problem* $Ax = b$ *and a floating point arithmetic* $\langle \hat{+}, \hat{-}, \hat{\times}, \hat{/} \rangle$ *such that the solution* $\hat{x}$ *computed by Gaussian elimination with partial or complete pivoting satisfies* $(A + \Delta A)\hat{x} = b$ *only for those matrices* ΔA *for which*

$$\frac{\| |\Delta A| |\hat{x}| \|}{\| |A| |\hat{x}| \|} \geq \lceil 19 \cdot 2^{n-2} - n - 8 \rceil u + O(u^2).$$

Therefore, the bound of Theorem 4.1 *is the best possible bound in the limit* $u \to 0$.

PROOF. See Appendix A for the proof, which employs a modification of Wilkinson's [1963] example. Q.E.D.

The next theorem uses Theorem 4.1 to get a bound on the backward error. We are especially interested in the effect of scaling on the error bound.

THEOREM 4.3. *Let* $\mathcal{I} = \{i\colon \ (|A||\hat{x}|)_i = 0\}$. *If* $d_i = 0$ *for* $i \in \mathcal{I}$ *and* $d_i \neq 0$ *for* $i \notin \mathcal{I}$, *then the backward error*

$$\eta \leq \chi(n)u \frac{\max |D^{-1}A||\hat{x}|}{\min\limits_{i \notin \mathcal{I}} (|D^{-1}A||\hat{x}|)_i} .$$

PROOF. Putting $z = |\hat{x}|$ in Theorem 4.1 gives

$$\begin{aligned} \|D_\varepsilon^{-1}(b - A\hat{x})\| = \|D_\varepsilon^{-1} \Delta A \hat{x}\| &\leq \| |D_\varepsilon^{-1} \Delta A| |\hat{x}| \| \\ &\leq \chi(n)u \| |D_\varepsilon^{-1} A| |\hat{x}| \|. \end{aligned} \qquad (4.7)$$

Multiplying this by ε and letting $\varepsilon \to 0$ gives

$$\max_{i \in \mathcal{I}} |(b - A\hat{x})_i| \leq \chi(n)u \max_{i \in \mathcal{I}} (|A||\hat{x}|)_i,$$

from which it follows that $(b - A\hat{x})_i = 0$ for $i \in \mathcal{I}$. Hence, from Theorem 3.2 we have that

$$\eta = \max_{i \notin I} \frac{|(b - A\hat{x})_i|}{(|A||\hat{x}|)_i} = \max_{i \notin I} \frac{(|D^{-1}(b - A\hat{x})|)_i}{(|D^{-1}A||\hat{x}|)_i} ,$$

and from (4.7) we see that

$$(|D^{-1}(b - A\hat{x})|)_i \leq \chi(n)u \quad \max \quad |D^{-1}A||\hat{x}|. \qquad \text{Q.E.D.}$$

By choosing

$$d_i = (|A||\hat{x}|)_i \tag{4.8}$$

the bound on the backward error is minimized, giving $\eta \leq \chi(n)u$. This suggests that a linear system should be scaled by dividing each row by its weighted ℓ_1 norm where the weights are the components of the computed solution. Unfortunately, (4.8) represents an implicit equation for the scale factors d_i because the computed solution $\hat{x}$ is a function $\xi(\Pi(D))$ of the scaling matrix D; that is, D must solve the equation

$$D = \mathrm{diag}(|A||\xi(\Pi(D))|), \tag{4.9}$$

for which a solution may not exist. The nature of this equation becomes more apparent by noting that it is equivalent to solving for a permutation P that satisfies

$$P = \Pi(\mathrm{diag}(|A||\xi(P)|)). \tag{4.10}$$

For if D satisfies (4.9), then $P = \Pi(D)$ satisfies (4.10); and if P satisfies (4.10), then $D = \mathrm{diag}(|A||\xi(P)|)$ satisfies (4.9). In principle we could determine the solution to (4.9), if there is one, by testing to see if any of the n! permutations P satisfy (4.10). In the cases where a solution exists, the backward error is bounded by $\chi(n)u$. One suspects that (4.10) almost always has a solution, and it is even conceivable that (4.10) always has a solution, at least whenever $\xi(P)$ is defined for all P. The existence of a solution for (4.10) implies the existence of an ordering for the rows which makes Gaussian elimination stable.

If one wishes to solve (4.10), the following iteration would likely converge and converge quickly for almost every system of equations:

$$P_{(0)} = \Pi(\mathrm{diag}(|A|e)),$$

$$P_{(m+1)} = \Pi(\mathrm{diag}(|A||\xi(P_{(m)})|)), \quad m = 1, 2, \ldots .$$

This is not suggested as a practical algorithm though because poorly scaled equations can usually be accurately solved by doing iterative improvement.

A more useful application of Theorem 4.3 is the diagnosis of ill-scaling, for

$$\sigma_R(A, \hat{x}) = \frac{\max|A||\hat{x}|}{\min|A||\hat{x}|}$$

is an easily computable measure of how badly scaled the rows are.

Remark. The bound of Theorem 4.1 can be refined:

$$(|D^{-1}\Delta A|z)_i \leq \chi(n)u \max_{j \leq i} (|D^{-1}A|z)_j .$$

This suggests that

$$\max_{\substack{i,j \\ i \geq j}} \frac{(|A||\hat{x}|)_j}{(|A||\hat{x}|)_i}$$

would be a better measure of the possible effect of ill scaling.

The quantity $\sigma_R(A, \hat{x})$ is not very satisfactory for theoretical purposes because $\hat{x}$ depends on the arithmetic used in the computation. We would prefer to use $\sigma_R(A, x)$ for the theory. For Hamming's example $|A||x| = (3 + 3\varepsilon, 6\varepsilon, 4\varepsilon)^T$ and $\sigma_R(A,x) = \frac{3}{4\varepsilon} + \frac{3}{4}$. Near optimal row scaling for this problem is given by

$$D^{-1}A = \begin{bmatrix} 3 & 2 & 1 \\ \frac{2}{\varepsilon} & 2 & 2 \\ \frac{1}{\varepsilon} & 2 & -1 \end{bmatrix}, \quad D^{-1}b = \begin{bmatrix} 3 + 3\varepsilon \\ 6 \\ 2 \end{bmatrix} .$$

For Gear's example $|A||x| = (2/\varepsilon + 2, 1, 1, 2)$ and $\sigma_R(A,x) = 2/\varepsilon + 2$. Near optimal row scaling is given by

$$D^{-1}A = \begin{bmatrix} \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ 0 & \varepsilon & 0 & 0 \\ 0 & 0 & \varepsilon & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad D^{-1}b = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 2 \end{bmatrix}.$$

One may wonder about the effect of scaling strategies such as row equilibration. Van der Sluis [1970b, p. 80] gives an example showing "that it is quite possible...that there exists no bound *depending on* n *only* for the ratios of the errors after and before equilibration." He goes on to describe a cautious equilibration scheme that never worsens the situation at the expense of possibly not improving it. An adaptation of this scheme to our theory is to choose

$$d_i = \min_k \max_j \left|\frac{a_{ij}}{a_{kj}}\right|,$$

which has the effect of leaving no row of A strictly dominated by any other row of A. Note that, since $d_i \le 1$, we have $\min|D^{-1}A||x| \ge A||x|$. Furthermore,

$$\begin{aligned}(|A||x|)_i &= \min_k \sum_\ell \left|\frac{a_{i\ell}}{a_{k\ell}}\right||a_{k\ell}||x_\ell| \\ &\le \min_k \max_j \left|\frac{a_{ij}}{a_{kj}}\right| \sum_\ell |a_{k\ell}||x_\ell| \\ &\le d_i \max|A||x|,\end{aligned}$$

whence $\max|D^{-1}A||x| \le \max|A||x|$. Therefore

$$\sigma_R(D^{-1}A, x) = \frac{\max|D^{-1}A||x|}{\min|D^{-1}A||x|} \le \frac{\max|A||x|}{\min|A||x|} = \sigma_R(A, x),$$

so that the scaling of the problem is not made worse by our choice of D.

The theorem that follows gives bounds on the backward error that in the limit u → 0 depend only on how ill scaled the problem is and not how ill conditioned it is. First we need a lemma.

LEMMA 4.4. *If* D *is nonsingular, then*

$$\|D^{-1}\Delta Ax\| \le \frac{\chi(n)u\,\|\,|D^{-1}A|\,|x|\,\|}{1-\chi(n)u\ \mathrm{Cond}(A^{-1}D)} .$$

PROOF. We have

$$\hat{x} + A^{-1}\Delta A\hat{x} = x,$$

which implies

$$D^{-1}\Delta A\hat{x} = (I + D^{-1}\Delta A\ {}^{-1}D)^{-1}D^{-1}\Delta Ax$$

and

$$\|D^{-1}\Delta Ax\| \le \frac{\|D^{-1}\Delta Ax\|}{1-\|D^{-1}\Delta A\ A^{-1}\ D\|} .$$

The term in the numerator

$$\begin{aligned}\|D^{-1}\Delta Ax\| &\le \|\,|D^{-1}\Delta A|\,|x|\,\| \\ &\le \chi(n)u\,\|\,|D^{-1}A|\,|x|\,\|,\end{aligned}$$

and the term in the denominator

$$\begin{aligned}\|D^{-1}\Delta A\ A^{-1}D\| &= \|\,|D^{-1}\Delta A\ A^{-1}D|e\| \\ &\le \|\,|D^{-1}\Delta A|\,|A^{-1}D|e\| \\ &\le \chi(n)u\,\|\,|D^{-1}A|\,|A^{-1}D|e\| \\ &= \chi(n)u\,\|\,|D^{-1}A|\,|A^{-1}D|\,\| .\end{aligned}$$

Q.E.D.

THEOREM 4.5. *Let* D *be nonsingular and let* $|A|\,|x| > 0$. *Gaussian elimination with row pivoting gives*

$$\eta \le \frac{\chi(n)u\ \sigma_R(D^{-1}A,\ x)}{1-2\chi(n)u\ \mathrm{Cond}(A^{-1}D)\ \sigma_R(D^{-1}A,\ x)}$$

provided that the denominator is positive.

PROOF. Choose ΔA as in Theorem 4.1. From Theorem 3.2 we have that

$$\eta = \max\frac{|-\Delta A\hat{x}|}{|A|\,|\hat{x}|} .$$

Since $x = \hat{x} + A^{-1}\, A\hat{x}$, we have

$$|x| \le |\hat{x}| + |A^{-1}D|e\|D^{-1}\Delta A\hat{x}\| ,$$

and so

$$|D^{-1}A||x| \le |D^{-1}A||\hat{x}| + |D^{-1}A||A^{-1}D|e\|D^{-1}\Delta Ax\| .$$

Thus

$$\eta \le \max \frac{e\|D^{-1}\Delta A\hat{x}\|}{|D^{-1}A||\hat{x}|}$$

$$\le \max \frac{e\|D^{-1}\Delta A\hat{x}\|}{|D^{-1}A||x| - e\ \mathrm{Cond}(A^{-1}D)\|D^{-1}\Delta A\hat{x}\|} .$$

Applying Lemma 4.4 yields

$$e\|D^{-1}\Delta A\hat{x}\| \le \frac{\chi(n)u\ \sigma_R(D^{-1}A,\ x)}{1 - \chi(n)u\ \mathrm{Cond}(A^{-1}D)}\ |D^{-1}A||x| ,$$

from which the theorem follows. Q.E.D.

Although we are unable to prove that there is always some ordering of the rows for which Gaussian elimination is stable, we can show that this is true asymptotically as $u \to 0$.

THEOREM 4.6. *For any problem such that* $|A||x| > 0$ *there is some ordering of the rows for which Gaussian elimination is asymptotically stable.*

PROOF. Using Theorem 4.5 with $D = \mathrm{diag}(|A||x|)$ gives

$$\eta \le \frac{\chi(n)u}{1 - 2\chi(n)u \max \dfrac{|A||A^{-1}||A||x|}{|A||x|}}$$

for small enough u. Hence for

$$u \le \bar{u}(n) = \min \frac{|A||x|}{4\chi(n)|A||A^{-1}||A||x|}$$

we have

$$\eta \le 2\chi(n)u. \qquad \text{Q.E.D.}$$

We end this section by examining the effect of scaling on a good bound for the "forward" error.

THEOREM 4.7. *The error*

$$\|\hat{x} - x\| \leq \frac{\chi(n)u\,\|A^{-1}D\|\,\|\,|D^{-1}A|\,|x|\,\|}{1 - \chi(n)u\,\mathrm{Cond}(A^{-1}D)} .$$

PROOF. We have

$$\begin{aligned}\|\hat{x} - x\| &= \|A^{-1}(-\,\Delta A\hat{x})\| \\ &\leq \|A^{-1}D\|\,\|D^{-1}\Delta A\hat{x}\| ,\end{aligned}$$

and the theorem follows from Lemma 4.4. Q.E.D.

If higher order terms in u are ignored, the bound on the error is minimized by choosing $D = \mathrm{diag}(|A|\,|x|)$. Thus

$$\frac{\|A^{-1}\|\,\|\,|A|\,|x|\,\|}{\|\,|A^{-1}|\,|A|\,|x|\,\|}$$

is a measure of the possible effect on the "forward" error of how poorly the equations happen to be scaled. For Hamming's example this quantity is $\frac{6}{17}\,\varepsilon^{-1} + O(1)$ and for Gear's it is $\varepsilon^{-1} + O(1)$.

The usual type of bound on the error is of the form

$$\|\hat{x} - x\| \leq \chi(n)u\|A^{-1}D\|\,\|D^{-1}A\|\,\|x\| + O(u^2).$$

This is minimized by $D = \mathrm{diag}(|A|e)$, which is row equilibration with the ℓ_1 norm.

5. *Gaussian Elimination with Column Pivoting and Scaling*

This section is similar to the previous section except that we examine the variant of Gaussian elimination in which the columns are interchanged in order to ensure that the pivot element is the largest in its row. The algorithm is assumed to do column scaling where the scale factors $d_1, d_2, \ldots, d_n$ are given as inputs to the algorithm. Again the selection of the pivot is assumed to be done exactly so that

$$\left| \frac{a_{ki}^{(k)} d_i}{a_{kk}^{(k)}} \right| \le |d_k|, \quad i \ge k+1 .$$

Writing the condition in this form allows for the use of zero scale factors but does not permit the selection of a zero pivot.

An *a priori* error bound is given by the following theorem:

THEOREM 5.1. *Let the vector* $\hat{x}$ *be computed by Gaussian elimination with column pivoting and column scaling where* $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$ *is the matrix of scale factors. Then there exists* ΔA *such that*

$$(A + \Delta A)\hat{x} = b$$

with $$|\Delta A D|e \le \tilde{\chi}(n)u|AD|e$$

where $$\tilde{\chi}(n) = \lfloor 27 \cdot 2^{n-2} - 5n - 7 \rfloor e^{2nu}.$$

PROOF. See Appendix B. Q.E.D.

It is likely that the constant $\tilde{\chi}(n)$ in this bound could be replaced by a smaller constant.

The following theorem indicates how the columns should be scaled in order that Gaussian elimination with column pivoting be stable.

THEOREM 5.2. *Let* $\hat{x}$ *be the value of* $A^{-1}b$ *computed by Gaussian elimination with column pivoting and column scaling where* $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$ *is the matrix of scale factors. Let* $I = \{i\colon (|A||\hat{x}|)_i = 0\}$ *and let* $J = \{j\colon \hat{x}_j = 0\}$. *If* $d_j = 0$ *for* $j \in J$ *and* $d_j \ne 0$ *for* $j \notin J$, *then the backward error*

$$\eta \leq \tilde{\chi}(n)u \max_{i \notin I} \frac{(|AD|e)_i}{(|A||\hat{x}|)_i} \max_{j \notin J} (|D^{-1}\hat{x}|)_j .$$

PROOF. We have $\hat{x} = DD_1^{-1}\hat{x}$ where D_1 denotes D with all diagonal zero entries replaced by ones. Hence

$$\begin{aligned} |b - A\hat{x}| &= |\Delta A\hat{x}| \leq |\Delta AD|e \; ||D_1^{-1}\hat{x}|| \\ &\leq \tilde{\chi}(n)u \; |AD|e \; ||D_1^{-1}\hat{x}|| \\ &= \tilde{\chi}(n)u \; |AD|e \max_{j \notin J} (|D^{-1}\hat{x}|)_j . \end{aligned}$$

We have that $(|A||\hat{x}|)_i = 0$ implies that $a_{ij} = 0$ for $j \notin J$, which implies that $(|AD|e)_i = 0$. Hence the theorem follows from Theorem 3.2. Q.E.D.

COROLLARY. *The backward error*

$$\eta \leq \tilde{\chi}(n)u \max_{i,j \notin J} \frac{(|D^{-1}\hat{x}|)_j}{(|D^{-1}\hat{x}|)_i} .$$

PROOF. We have

$$e \leq |D^{-1}\hat{x}| \max_{j \notin J} \frac{1}{(|D^{-1}\hat{x}|)_j} .$$ Q.E.D.

A choice of D which minimizes the bound on the backward error is

$d_i = \hat{x}_i$;

that is, we scale by multiplying the i-th column by the i-th component of the computed solution. Again these weights are not known at the time when scaling is performed. The main value of this theorem is that it gives an easily computable measure of column ill scaling:

$$\sigma_C(A, \hat{x}) = \max \frac{|A|e \, ||\hat{x}||}{|A||\hat{x}|} .$$

For theoretical purposes we would prefer to use $\sigma_C(A, x)$. For Hamming's example $\sigma_C(A, x) = \frac{1}{3\varepsilon} + \frac{2}{3}$. Near optimal row and column scaling, which would be appropriate for complete pivoting, is given by

$$D_1^{-1}AD_2 = \begin{bmatrix} 3\varepsilon & 2 & 1 \\ 2 & 2 & 2 \\ 1 & 2 & -1 \end{bmatrix}, \qquad D_1^{-1}b = \begin{bmatrix} 3 + 3\varepsilon \\ 6 \\ 2 \end{bmatrix}.$$

For Gear's example $\sigma_C(A,\ x) = \frac{1}{\varepsilon}$. Near optimal row and column scaling is given by

$$D_1^{-1}AD_2 = \begin{bmatrix} \varepsilon & 1 & -1 & \varepsilon \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \qquad D_1^{-1}b = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 2 \end{bmatrix}.$$

LEMMA 5.3. *If* D *is nonsingular, then*

$$|\Delta A\hat{x}| \le \frac{\tilde{\chi}(n)u\ |AD|e\ \|D^{-1}x\|}{1 - \tilde{\chi}(n)u\ \mathrm{Cond}(AD)}$$

provided that the denominator is positive.

PROOF. We have

$$\hat{x} + A^{-1}\Delta A\hat{x} = x,$$

which implies

$$\Delta A\hat{x} = \Delta AD(I + D^{-1}A^{-1}\Delta AD)^{-1}\ D^{-1}x.$$

Hence,

$$|\Delta A\hat{x}| \le |\Delta AD|e\ \|(I + D^{-1}A^{-1}\Delta AD)^{-1}\ D^{-1}x\|$$

$$\le \frac{\tilde{\chi}(n)u\ |AD|e\ \|D^{-1}x\|}{1 - \|D^{-1}A^{-1}\Delta AD\|}\ .$$

The term in the denominator

$$\|D^{-1}A^{-1}\Delta AD\| \le \|\,|D^{-1}A^{-1}|\,|\Delta AD|e\|$$

$$\le \tilde{\chi}(n)u\ \|\,|D^{-1}A^{-1}|\,|AD|e\|$$

$$= \tilde{\chi}(n)u\ \|\,|D^{-1}A^{-1}|\,|AD|\,\|\ .$$

Q.E.D.

THEOREM 5.4. *Let* D *be nonsingular and let* $|A||x| > 0$. *Gaussian elimination with column pivoting gives*

$$\eta \leq \frac{\tilde{\chi}(n)u\ \sigma_C(AD,\ D^{-1}x)}{1 - 2\tilde{\chi}(n)u\ \mathrm{Cond}(AD)\ \sigma_C(AD,\ D^{-1}x)}$$

provided that the denominator is positive.

PROOF. Choose ΔA as in Theorem 5.1. From Theorem 3.2 we have that

$$\eta = \max \frac{|-\ \Delta A\hat{x}|}{|A||\hat{x}|}\ .$$

Furthermore

$$x = \hat{x} + A^{-1}\Delta A\hat{x},$$

and so

$$|A||\hat{x}| \geq |A||x| - |A||A^{-1}||\Delta A\hat{x}|.$$

Therefore

$$\eta \leq \max \frac{|\Delta A\hat{x}|}{|A||x| - |A||A^{-1}||\Delta A\hat{x}|}\ .$$

Applying Lemma 5.3 yields

$$|\Delta A\hat{x}| \leq \frac{\tilde{\chi}(n)u\ \sigma_C(AD,\ D^{-1}x)}{1 - \tilde{\chi}(n)u\ \mathrm{Cond}(AD)}\ |A||x|,$$

from which the theorem follows. Q.E.D.

THEOREM 5.5. *For any problem such that* $|x| > 0$ *there is some ordering of the columns for which Gaussian elimination is asymptotically stable.*

PROOF. Using Theorem 5.4 with $D = \mathrm{diag}(|x|)$ gives

$$\eta \leq \frac{\tilde{\chi}(n)u}{1 - 2\tilde{\chi}(n)u\ \max \frac{|A^{-1}||A||x|}{|x|}}$$

for small enough u. Hence for

$$u \leq \bar{u}(n) = \min \frac{|x|}{4\tilde{\chi}(n)|A^{-1}||A||x|}$$

we have

$$\eta \leq 2\tilde{\chi}(n)u.$$ Q.E.D.

Recall that the stability threshold in the case of Gaussian elimination with optimal row ordering was

$$\tilde{u}(n) = \min \frac{|A||x|}{4\tilde{\chi}(n)|A||A^{-1}||A||x|} .$$

It is easy to show that this is larger than the stability threshold for optimal column ordering. This is a slight indication that row pivoting may be superior to column pivoting.

THEOREM 5.6. *The error*

$$||\hat{x} - x|| \leq \frac{\tilde{\chi}(n)u\ ||\,|A^{-1}|\,|AD|\,||\ ||D^{-1}x||}{1 - \tilde{\chi}(n)u\ \mathrm{Cond}(AD)} .$$

PROOF. Since

$$(A + \Delta A)(\hat{x} - x) = - \Delta Ax,$$

we have

$$\begin{aligned} \hat{x} - x &= -(I + A^{-1}\Delta A)^{-1}\ A^{-1}\Delta Ax \\ &= -\ A^{-1}\Delta A(I + A^{-1}\Delta A)^{-1}x \\ &= -\ A^{-1}\Delta AD(I + D^{-1}A^{-1}\Delta AD)^{-1}\ D^{-1}x. \end{aligned}$$

Therefore,

$$\begin{aligned} ||\hat{x} - x|| &\leq \frac{||A^{-1}\ AD||\ ||D^{-1}x||}{1 - ||D^{-1}A^{-1}\Delta AD||} \\ &\leq \frac{|A^{-1}|\,|\Delta AD|e\ ||D^{-1}x||}{1 - ||\,|D^{-1}A^{-1}|\,|\Delta AD|e||} \\ &\leq \frac{\tilde{\chi}(n)u\ ||\,|A^{-1}|\,|AD|\,||\ ||D^{-1}x||}{1 - \tilde{\chi}(n)u\ \mathrm{Cond}(AD)} . \end{aligned}$$

Q.E.D.

If higher order terms in u are ignored, the bound on the error is minimized by choosing D = diag(|x|). Thus for column pivoting

$$\frac{\| |A^{-1}| |A| \| \|x\|}{\| |A^{-1}| |A| |x| \|}$$

is a measure of the possible effect of ill scaling on the "forward" error.

6. *Iterative Improvement*

It is often thought that iterative improvement is not worthwhile unless either (i) the uncertainty in the values of A is less than the unit roundoff error (e.g., if the elements of A are integers) or (ii) we wish to diagnose ill conditioning. This thinking is based on the fact that Gaussian elimination with pivoting is stable from the absolute error point of view. But according to the relative error point of view, Gaussian elimination may not give acceptable accuracy, and so it is of interest to examine the stability behavior of iterative improvement. Results of a careful error analysis are given for iterative improvement both with and without double precision accumulation of the residuals.

The algorithm being considered is described as follows where subscripts denote iterates rather than components of vectors:

x_1 = value of $A^{-1}b$ computed by row pivoting,

for m = 1, 2, 3,...

r_m = computed value of $b - Ax_m$,

d_m = value of $A^{-1}r_m$ computed by row pivoting,

$x_{m+1} = x_m \hat{+} d_m$.

The algorithm for computing r_m appears in §3 and the algorithm for d_m is in §4.

The theorem which follows shows that just one iteration of iterative improvement with just single precision accumulation of the residuals is enough to make Gaussian elimination asymptotically stable. This may seem to contradict the usual advice (Forsythe and Moler [1967, p. 49]) that "It is absolutely essential that the residual r_k be computed with a higher precision than that of the rest of the computation." Actually there is little conflict because we have shown that poorly scaled systems may be solved with an effective precision of much less than single precision.

THEOREM 6.1. *Assume that* $|A||x| > 0$. *Gaussian elimination followed by one iteration of iterative improvement results in a backward error which satisfies*

$$\eta_2' \leq (n+1)u + \{(\chi(n)^2 + 2n\chi(n) + n^2 + 2n)\ \mathrm{Cond}(A^{-1})\ \sigma_R(A, x) + \chi(n)\ \sigma_R(A, x) + \frac{1}{2}n^2 + \frac{1}{2}n\}u^2 + O(u^3)$$

for single precision residual accumulation and

$$\eta_2'' \leq u + \{\chi(n)^2\ \mathrm{Cond}(A^{-1})\ \sigma_R(A, x) + \chi(n)\ \sigma_R(A, x) + n\}u^2 + O(u^3)$$

for double precision residual accumulation. That is, the algorithm is asymptotically stable in either case.

PROOF. The theorem follows from Theorem C.9 in Appendix C. Q.E.D.

Note. The asymptotic nature of these bounds conceals the fact that certain assumptions on the smallness of u are necessary in order to get any bound at all. The actual assumptions, found in Appendix C, are too lengthy to reproduce here; roughly speaking it must be assumed that the coefficient of the second order term is less than 1/u.

Recall from Theorem 4.3 that for no iterations we have

$$\eta_1 \leq \chi(n)\ \sigma_R(A, x)u + O(u^2),$$

and thus one iteration does make a big difference in the size of the bound. However, the presence of the product $\mathrm{Cond}(A^{-1})\ \sigma_R(A, x)$ in the second order term indicates that this may not be the case for problems which are sufficiently ill scaled and ill conditioned.

THEOREM 6.2. *Assume* $|A||x| > 0$. *The backward error* η_m *for iterative improvement of Gaussian elimination with row pivoting satisfies*

$$\overline{\lim_{m\to\infty}}\ \eta_m' \leq (n+1)u + \{2n(\chi(n) + n+1)\ \mathrm{Cond}(A^{-1})\ \sigma_R(A, x) + \chi(n)\ \sigma_R(A, x) + \frac{1}{2}n^2 + \frac{5}{2}n + 1\}u^2 + O(u^3)$$

for single precision residual accumulation and

$$\overline{\lim_{m\to\infty}}\ \eta_m'' \leq u + \{\chi(n)\ \sigma_R(A, x) + n+1\}u^2 + O(u^3)$$

for double precision residual accumulation.

PROOF. The theorem follows from Theorem C.7 in Appendix C. Q.E.D.

The main effect of doing more iterations with single precision accumulations is a moderate reduction in the magnitude of the second term. But for the double precision case there is a striking improvement due to the disappearance of the $\chi(n)^2$ Cond(A^{-1}) $\sigma_R(A, x)u^2$ term so that the bound on η''_m depends on the condition number of A^{-1} only through the $O(u^3)$ term. This may represent a significant improvement for problems which are both poorly scaled and ill conditioned.

7. *Practical Implications*

The comments that follow are suggested by the error analysis, but their usefulness remains to be established.

By means of examples it has been shown that Gaussian elimination with (partial or complete) pivoting does not generally provide all the accuracy that the data deserve or even a fixed fraction of that accuracy. Hamming [1971, p. 121] states

> "It is reasonable to ask how typical these examples are and how often in the past the pivoting method has created the ill conditioning that was reported to occur by some library routines. The answers are not known at this time; all that is claimed is that textbooks and library descriptions rarely, if ever, mention this possibility (though it is apparently known in the folklore)."

and so it seems that there have been practical instances where the pivoting method has performed poorly. Perhaps Gaussian elimination *without* iterative improvement should be regarded as a "quick and dirty" way to solve linear equations.

The computation of the backward error is one reliable test for deciding whether or not the solution of a linear system is "reasonably accurate." The test can be made quite efficient by accumulating r and $|A||\hat{x}| + |b|$ at the same time. If the test is failed, then in most cases the use of iterative improvement would result in a solution which passes the test. One could, of course, forgo the backward error computation and just do iterative improvement until "convergence." But such a procedure may not be completely reliable since it has not been rigorously proven that "convergence" implies a reasonably accurate solution. Stewart [1973, p. 205] mentions "the possibility that, with a violently ill-conditioned matrix, the iteration may appear to converge to a false solution."

It is also suggested by the theory that if double precision accumulation of the residuals is costly, then iterative improvement with single precision accumulation might still be beneficial.

The success of the pivoting method depends upon a reasonable scaling of the equations, which is at best guesswork unless one has some knowledge about the sizes of the solution components. If $c \simeq |x|$, then

(i) for row pivoting one should scale the system to get

$(D_1^{-1}A)x = (D_1^{-1}b)$ where $D_1 = \mathrm{diag}(|A|c)$.

(ii) for complete pivoting one should scale the system to get

$(D_1^{-1}AD_2)(D_2^{-1}x) = (D_1^{-1}b)$ where $D_1 = \mathrm{diag}(|A|c)$ and $D_2 = \mathrm{diag}(c)$.

It may be worthwhile to allow users of a linear equation solver to provide an estimate of the solution, particularly if the solution variable X is being used only as an output parameter. For simple use of the program X could be set to all ones.

Appendix A. Error Bounds for Row Pivoting

For Lemmas A.1 through A.8 it is assumed that D has nonzero diagonal elements. This assumption does not apply to the theorem that follows these lemmas. For any $n \times n$ matrix $C = (c_{ij})$ let $\bar{c}_{ij} = c_{ij}/d_i$. Also, let $\omega = 1 + u$.

LEMMA A.1. *We have*

$$|m_{ik}| \le \omega |d_i/d_k|, \qquad i > k,$$

and

$$|\bar{a}_{ij}^{(k)}| \le \omega^{k+1} \sum_{\ell=1}^{k-1} (2\omega)^{k-1-\ell} |\bar{a}_{\ell j}| + \omega^{k-1} |\bar{a}_{ij}|, \qquad i,j \ge k.$$

PROOF. Equation (4.1) implies

$$|m_{ik}| \le \omega |a_{ik}^{(k)}/a_{kk}^{(k)}|, \qquad i > k,$$

and because of row pivoting (see (4.5)) we get the first inequality of the lemma. Equation (4.2) implies

$$|a_{ij}^{(k+1)}| \le \omega |a_{ij}^{(k)}| + (1 + 2u)|m_{ik}\, a_{kj}^{(k)}|, \qquad i,j \ge k + 1$$

and therefore

$$|\bar{a}_{ij}^{(k+1)}| \le \omega |\bar{a}_{ij}^{(k)}| + \omega(1 + 2u)|\bar{a}_{kj}^{(k)}|, \qquad i,j \ge k + 1.$$

The second inequality of the lemma follows from this by induction on k. Q.E.D.

LEMMA A.2. *The matrices* $L = (\ell_{ij})$ *and* $U = (u_{ij})$ *satisfy*

$$LU = A + E^{(1)} + E^{(2)} + \ldots + E^{(n-1)}$$

where the matrices $E^{(k)}$ *have elements* $\varepsilon_{ij}^{(k)}$ *which satisfy*

$$|\varepsilon_{ij}^{(k)}| \le \begin{cases} \omega^{-1} u |a_{ij}^{(k)}| + (2 + 3u)\, \omega^{-2} u |m_{ik} a_{kj}^{(k)}| & \textit{for } i,j > k\,, \\ \omega^{-1} u |a_{ik}^{(k)}| & \textit{for } i > j = k, \\ 0 & \textit{otherwise,} \end{cases}$$

regardless of the pivoting strategy.

PROOF. Define the elements of $E^{(k)}$ by

$$\varepsilon_{ij}^{(k)} = \begin{cases} a_{ij}^{(k+1)} - a_{ij}^{(k)} + m_{ik}a_{kj}^{(k)} & \text{for } i,j > k, \\ m_{ik}a_{kk}^{(k)} - a_{ik}^{(k)} & \text{for } i > j = k, \\ 0 & \text{otherwise.} \end{cases}$$

By separately considering the cases $i \leq j$ and $i > j$ it is straightforward to show that the elements $\varepsilon_{ij}^{(k)}$ of $E^{(k)}$ satisfy

$$\sum_{k=1}^{n-1} \varepsilon_{ij}^{(k)} = \sum_{k=1}^{n} \ell_{ik} u_{kj} - a_{ij},$$

which establishes the equality of the lemma. Let $k \leq n - 1$ be fixed. Write (4.1) as

$$m_{ik} = (a_{ik}^{(k)}/a_{kk}^{(k)})(1 + \delta_{ik}), \qquad i \geq k + 1,$$

and (4.2) as

$$a_{ij}^{(k+1)} = (a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)}(1 + \delta_{ij}))(1 + \delta'_{ij}), \; i,j \geq k + 1,$$

where the δ's are relative roundoff errors. Then

$$\varepsilon_{ij}^{(k)} = \begin{cases} a_{ij}^{(k)}\delta'_{ij} - m_{ik}\, a_{kj}^{(k)}(\delta_{ij} + \delta'_{ij} + \delta_{ij}\, \delta'_{ij}) & \text{for } i,j > k, \\ a_{ik}^{(k)}\delta_{ik} & \text{for } i > j = k, \\ 0 & \text{otherwise,} \end{cases}$$

and the lemma follows from the bound $u/(1 + u)$ on the δ's. Q.E.D.

LEMMA A.3. *The matrices* L *and* U *satisfy*

$$LU = A + E$$

with

$$\| |D^{-1}E|z\| \leq (3 \cdot 2^{n-1} - 3)\omega^{2n-1}u\| |D^{-1}A|z\|$$

for arbitrary $z \geq 0$.

PROOF. Let $E = E^{(1)} + E^{(2)} + \ldots + E^{(n-1)}$ where the $E^{(k)}$ are given by Lemma A.2. Substituting the bound on m_{ik} of Lemma A.1 into the bound on the elements $\varepsilon_{ij}^{(k)}$ we get

$$|\bar{\varepsilon}_{ij}^{(k)}| \le \omega^{-1}u|\bar{a}_{ij}^{(k)}| + (2 + 3u)\omega^{-1}u|\bar{a}_{kj}^{(k)}|,$$

which, in fact, is valid for all i,j. This implies

$$\sum_j |\bar{\varepsilon}_{ij}^{(k)}|z_j \le 3u\| |D^{-1}A^{(k)}|z\|$$

for arbitrary $z \ge 0$. From Lemma A.1 it immediately follows that $\| |D^{-1}A^{(k)}|z\| \le 2^{k-1}\omega^{2k-1}\| |D^{-1}A|z\|$; and so we have

$$\| |D^{-1}E^{(k)}|z\| \le 3\cdot 2^{k-1}\omega^{2n-1}u\| |D^{-1}A|z\|,$$

from which the lemma follows. Q.E.D.

LEMMA A.4. *We have*

$$|\ell_{ij}| \le \omega|d_i/d_j|, \qquad i > j,$$

and

$$|\bar{u}_{ij}| \le \omega^{2i-1} \sum_{\ell=1}^{i} \lceil 2^{i-\ell-1}\rceil |\bar{a}_{ij}|, \qquad i \le j.$$

PROOF. Follows from Lemma A.1 because $\ell_{ij} = m_{ij}$ and $u_{ij} = a_{ij}^{(i)}$. Q.E.D.

LEMMA A.5. *The vector* $\hat{x}$ *computed by* (4.4) *satisfies* $(U + \delta U)\hat{x} = y$ *for some upper triangular matrix* δU *such that*

$$|\delta u_{ij}| \le g(i,j)\omega^{n-i}|u_{ij}| \text{ and } |u_{ij} + \delta u_{ij}| \le \omega^{n-i+1}|u_{ij}|$$

where

$$g(i,j) = \begin{cases} n - j + 2, & j \ge i + 2, \\ n - j + 1, & j = i + 1, \\ 2, & j = i \le n - 1, \\ 1, & j = i = n, \end{cases}$$

regardless of the pivoting strategy.

PROOF. From (4.4) we have that

$$\frac{u_{ii}\hat{x}_i}{(1+\delta_i)(1+\delta'_i)} + (\dots(u_{i,i+1} \mathbin{\hat{\times}} \hat{x}_{i+1} \mathbin{\hat{+}} u_{i,i+2} \mathbin{\hat{\times}} \hat{x}_{i+2})\dots \mathbin{\hat{+}} u_{in} \mathbin{\hat{\times}} \hat{x}_n) = y_i, \quad i \le n-1,$$

and

$$\frac{u_{nn}\hat{x}_n}{1+\delta'_n} = y_n,$$

where δ_i and δ'_i are relative roundoff errors due to subtraction and division, respectively. By the usual type of analysis we can obtain the bounds

$$|\delta u_{ij}| \le \begin{cases} (\omega^{n-j+2} - 1)|u_{ij}|, & j \ge i+2, \\ (\omega^{n-j+1} - 1)|u_{ij}|, & j = i+1, \\ (\omega^{2} - 1)|u_{ij}|, & j = i \le n-1, \\ (\omega - 1)|u_{ij}|, & j = i = n, \end{cases}$$

from which the lemma follows. Q.E.D.

LEMMA A.6. *The matrix* δU *of Lemma* A.5 *satisfies*

$$\| |D^{-1}L\delta U|z\| \le \lceil 5\cdot 2^{n-2} - 2\rceil\, \omega^{2n} u \| |D^{-1}A|z\|$$

for arbitrary $z \ge 0$.

PROOF. From Lemma A.4 it follows that

$$(|D^{-1}L\delta U|z)_i \le \sum_j \sum_k |d_i^{-1}\ell_{ik}d_k|\,|\delta\bar{u}_{kj}|z_j$$

$$\le \omega \sum_j \sum_k |\delta\bar{u}_{kj}|z_j.$$

Applying Lemma A.5 first and then Lemma A.4 gives

$$(|D^{-1}L\delta U|z)_i \le u \sum_j \sum_k g(k,j)\omega^{n-k+1}|\bar{u}_{kj}|z_j$$

$$\le u \sum_{j=1}^{n} \sum_{k=1}^{j} g(k,j)\omega^{n+k} \sum_{\ell=1}^{k} \lceil 2^{k-\ell-1}\rceil |\bar{a}_{\ell j}|z_j$$

$$\le \omega^{2n} u \sum_{\ell=1}^{n} \sum_{j=\ell}^{n} \sum_{k=\ell}^{j} \lceil 2^{k-\ell-1}\rceil g(k,j)|\bar{a}_{\ell j}|z_j.$$

After much manipulation it turns out that

$$\max_{\ell \le j \le n} \sum_{k=\ell}^{j} \lceil 2^{k-\ell-1} \rceil \, g(k,j) = \lfloor 5 \cdot 2^{n-\ell-2} \rfloor ,$$

and therefore

$$(|D^{-1} L \delta U| z)_i \le \omega^{2n} u \sum_{\ell=1}^{n} \lfloor 5 \cdot 2^{n-\ell-2} \rfloor \, (|D^{-1} A| z)_\ell ,$$

from which the lemma follows. Q.E.D.

LEMMA A.7. *The vector* y *computed by* (4.3) *satisfies*

$$(L + \delta L) y = b$$

for some lower triangular matrix δL *whose elements* $\delta\ell_{ij}$ *satisfy*

$$|\delta\ell_{ij}| \le \min\,\{n - j + 1,\ n - 1\} \omega^{n-j} u |a_{ij}^{(j)} / a_{jj}^{(j)}| , \qquad i \ge j,$$

regardless of the pivoting strategy.

PROOF. We have from (4.3) that

$$(\ldots(\ell_{i1} \mathbin{\hat{\times}} y_1 \mathbin{\hat{+}} \ell_{i2} \mathbin{\hat{\times}} y_2)\ldots \mathbin{\hat{+}} \ell_{i,i-1} \mathbin{\hat{\times}} y_{i-1}) + \frac{y_i}{1 + \delta_i} = b_i , \qquad i \ge 2,$$

where δ_i is a relative roundoff error. By the usual type of analysis we get that

$$|\delta\ell_{i1}| \le ((1 + \omega^{-1} u)^{i-1} - 1) |\ell_{i1} , \qquad i \ge 2,$$
$$|\delta\ell_{ij}| \le ((1 + \omega^{-1} u)^{i-j+1} - 1) |\ell_{ij}| , \qquad i > j \ge 2,$$

and

$$|\delta\ell_{ii}| \le u |\ell_{ii}| .$$

The lemma follows from the inequalities

$$(1 + \omega^{-1} u)^k - 1 \le \omega^{-1} u \, k (1 + \omega^{-1} u)^{k-1}$$
$$\le k \omega^{k-2} u$$

and

$$|\ell_{ij}| \le \min\,\{\omega,\ \omega^{i-j}\} |a_{ij}^{(j)} / a_{jj}^{(j)}| . \quad \text{Q.E.D.}$$

LEMMA A.8. *The matrices* δU *and* δL *of Lemmas* A.5 *and* A.7 *satisfy*

$$\| |D^{-1}\delta L(U + \delta U)|z\| \leq (2^{n+1} - n - 3)\omega^{2n}u\| |D^{-1}A|z\|$$

for arbitrary $z \geq 0$.

PROOF. Using the bound of Lemma A.7 and the row pivoting inequality (4.5), we get

$$|\delta\ell_{ij}| \leq \min\{n - j + 1,\ n - 1\}\omega^{n-j}u|d_i/d_j|, \qquad i \geq j.$$

Hence

$$(|D^{-1}\delta L(U + \delta U)|z)_i \leq u \sum_j \min\{n - j + 1,\ n - 1\}\omega^{n-j}(|D^{-1}(U + \delta U)|z)_j.$$

It immediately follows from Lemma A.5 that

$$(|D^{-1}(U + \delta U)|z)_j \leq \omega^{n-j+1}(|D^{-1}U|z)_j$$

and from Lemma A.4 that

$$(|D^{-1}(U + \delta U)|z)_j \leq \omega^{n+j}\ 2^{j-1}\| |D^{-1}A|z\|.$$

Therefore,

$$(|D^{-1}\delta L(U + \delta U)|z)_i \leq \omega^{2n}u \sum_{j=1}^{n} \min\{n - j + 1,\ n - 1\}2^{j-1}\| |D^{-1}A|z\|,$$

from which the lemma follows. Q.E.D.

THEOREM 4.1. *Let the vector* $\hat{x}$ *be computed by Gaussian elimination with row pivoting and row scaling where* $D = \text{diag}(d_1, d_2, \ldots, d_n)$ *is the matrix of reciprocal scale factors. Then there exists* ΔA *such that* $(A + \Delta A)\hat{x} = b$ *with*

$$\| |D_\varepsilon^{-1}\Delta A|z\| \leq \chi(n)u\| |D_\varepsilon^{-1}A|z\|$$

for arbitrary $z \geq 0$ *and arbitrary* ε *satisfying* $0 < |\varepsilon| < \bar{\varepsilon}$ *where* $\chi(n) = \lceil 19\cdot 2^{n-2} - n - 8\rceil e^{2nu}$.

PROOF. The restriction on ε implies that $\xi(\Pi(D_\varepsilon)) = \xi(\Pi(D))$, and from Lemmas A.3, A.6, and A.8 we have the bounds

$$\| |D_\varepsilon^{-1}E|z\| \le (3\cdot 2^{n-1} - 3)\omega^{2n-1}u\| |D_\varepsilon^{-1}A|z\|,$$

$$\| |D_\varepsilon^{-1}L\delta U|z\| \le \lceil 5\cdot 2^{n-2} - 2\rceil\omega^{2n}u\| |D_\varepsilon^{-1}A|z\|,$$

and

$$\| |D_\varepsilon^{-1}\delta L(U + \delta U)|z\| \le (2^{n+1} - n - 3)\omega^{2n}u\| |D_\varepsilon^{-1}A|z\|.$$

The theorem follows from the equation

$$(A + E + \delta LU + (L + \delta L)\delta U)\hat{x} = b. \quad \text{Q.E.D.}$$

THEOREM 4.2. *There exists a problem* $Ax = b$ *and a floating point arithmetic* $<\hat{+}, \hat{-}, \hat{\times}, \hat{/}>$ *such that the solution* $\hat{x}$ *computed by Gaussian elimination with partial or complete pivoting satisfies* $(A + \Delta A)x = b$ *only for those matrices* ΔA *for which*

$$\frac{\| |\Delta A|\,|\hat{x}| \|}{\| |A|\,|x| \|} \ge \lceil 19\cdot 2^{n-2} - n - 8\rceil u + O(u^2).$$

Therefore, the bound of Theorem 4.1 *is the best possible bound in the limit* $u \to 0$.

PROOF. Obvious for $n = 1$. Assume $n \ge 2$. Let

$$A = \begin{bmatrix} M & & & & 1 \\ -M & M & & \bigcirc & 1 \\ \vdots & \vdots & \ddots & & \vdots \\ & & & M & 1 \\ -M & -M & \cdots & -M & 1 \end{bmatrix}$$

and

$$b_k = \begin{cases} 1 + (7\cdot 2^{k-1} - k - 8)u, & k \le n - 2, \\ 1 + (2^{n+1} - n - 7)u, & k = n - 1, \\ 1 + (19\cdot 2^{n-2} - n - 8)u, & k = n. \end{cases}$$

If M is large enough, then there are no interchanges even with complete pivoting. We have

$$m_{ij} = m = M\,\hat{/}\,(-M), \qquad i > j\ ,$$

$$a_{in}^{(1)} = 1,$$

$$a_{in}^{(k)} = a_{in}^{(k-1)} \,\hat{-}\, m \,\hat{\times}\, a_{k-1,n,}^{(k-1)}, \qquad i \ge k \ge 2.$$

Suppose that all these floating point operations increase the magnitude of the result by a factor $(1 + 2u)/(1 + u)$. Then

$$m = -(1 + u) + O(u^2)$$

and

$$a_{in}^{(k)} = (1 + u)\ a_{in}^{(k-1)} + (1 + 3u)\ a_{k-1,n}^{(k-1)} + O(u^2).$$

By induction on k it follows that

$$a_{in}^{(k)} = 2^{k-1} + 2^k(k - 1)u + O(u^2), \qquad i \geq k,$$

and hence

$$u_{kn} = 2^{k-1} + 2^k(k-1)u + O(u^2).$$

We have

$$y_1 = b_1$$

$$y_k = b_k \hat{-} S_{k-1}, \qquad k \geq 2,$$

where

$$S_k = (\ldots(m \hat{\times} y_1 \hat{+} m \hat{\times} y_2)\ldots \hat{+} m \hat{\times} y_k).$$

Then

$$S_1 = m \hat{\times} b_1$$

and

$$S_k = S_{k-1} \hat{+} m \hat{\times} (b_k \hat{-} S_{k-1}), \qquad 2 \leq k \leq n - 1.$$

Suppose that all these floating point operations reduce the magnitude of the result by a factor $1/(1 + u)$. Hence

$$S_1 = -b_1 + O(u^2)$$

and

$$S_k = (2 - 3u)S_{k-1} - (1 - 2u)b_k + O(u^2), \qquad 2 \leq k \leq n - 1.$$

By induction on k it follows that

$$S_k = 1 - 2^k - 2^{k+1}(k - 4)u - (k + 9)u + O(u^2), \qquad k \leq n - 2$$

and

$$S_{n-1} = 1 - 2^{n-1} - 2^{n-2}(4n - 19)u - (n + 8)u + O(u^2).$$

Also we have

$$y_1 = 1 - 2u,$$

$$y_k = (b_k - S_{k-1})(1 - u + O(u^2)), \qquad k \geq 2.$$

From this we get

$$y_k = 2^{k-1} + 2^k(k - 2)u + O(u^2), \qquad k \leq n - 2,$$

$$y_{n-1} = 2^{n-2} + 2^{n-2}(2n - 5)u + O(u^2),$$

$$y_n = 2^{n-1} + 2^{n-1}(2n - 1)u + O(u^2).$$

We have

$$\hat{x}_n = y_n \hat{/} u_{nn},$$

$$\hat{x}_{n-1} = (y_{n-1} \hat{-} u_{n-1,n} \hat{\times} \hat{x}_n) \hat{/} M,$$

$$\hat{x}_k = (y_k \hat{-} (0 \hat{+} u_{kn} \hat{\times} \hat{x}_n)) \hat{/} M, \qquad k \leq n - 2.$$

Suppose that all these floating point operations reduce the magnitude of the result by a factor $1/(1 + u)$. Then

$$\hat{x}_n = (y_n/u_{nn})(1 - u) = 1 + O(u^2),$$

$$u_{n-1,n} \hat{\times} \hat{x}_n = u_{n-1,n}(1 - u) = 2^{n-2} + 2^{n-2}(2n - 5)u + O(u^2),$$

$$\hat{x}_{n-1} = O(u^2),$$

$$0 \hat{+} u_{k,n} \hat{\times} \hat{x}_n = u_{k,n}(1 - 2u) = 2^{k-1} + 2^k(k - 2)u + O(u^2), \qquad k \leq n - 2,$$

and

$$\hat{x}_k = O(u^2), \qquad k \leq n - 2.$$

Hence

$$(b - A\hat{x})_n = b_n - 1 + O(u^2) = (19 \cdot 2^{n-2} - n - 8)u + O(u^2).$$

No matter how we choose ΔA we get

$$\| |\Delta A| |\hat{x}| \| \geq \|\Delta A\hat{x}\| = \|b - A\hat{x}\| = (19 \cdot 2^{n-2} - n - 8)u + O(u^2). \quad \text{Q.E.D.}$$

Appendix B. Error Bounds for Column Pivoting

For any matrix $C = (c_{ij})$ let $\bar{c}_{ij} = c_{ij}d_j$. Also, let $\omega = 1 + u$.

LEMMA B.1. *We have*

$$|m_{ik}\bar{a}_{kj}^{(k)}| \le \omega|\bar{a}_{ik}^{(k)}|, \qquad i > k,\ j \ge k,$$

and

$$|\bar{a}_{ij}^{(k)}| \le \omega^{k+1} \sum_{\ell=1}^{k-1} (2\omega)^{k-1-\ell}|\bar{a}_{i\ell}| + \omega^{k-1}|\bar{a}_{ij}|, \qquad i,j \ge k\ .$$

PROOF. Equation (4.1) implies

$$|m_{ik}a_{kj}^{(k)}d_j| \le \omega|a_{ik}^{(k)}a_{kj}^{(k)}d_j/a_{kk}^{(k)}|, \qquad i > k,\ j \ge k,$$

and because of (4.5) we get

$$|m_{ik}a_{kj}^{(k)}d_j| \le \omega|a_{ik}^{(k)}d_k|, \qquad i > k,\ j \ge k,$$

which proves the first inequality. Equation (4.2) implies

$$|a_{ij}^{(k+1)}| \le \omega|a_{ij}^{(k)}| + (1 + 2u)|m_{ik}a_{kj}^{(k)}|, \qquad i,j \ge k + 1,$$

and therefore

$$|\bar{a}_{ij}^{(k+1)}| \le \omega|\bar{a}_{ij}^{(k)}| + \omega(1 + 2u)|\bar{a}_{ik}^{(k)}|, \qquad i,j \ge k + 1.$$

The second inequality of the theorem follows from this by induction on k. Q.E.D.

LEMMA B.2. *The matrices* L *and* U *satisfy*

$$LU = A + E$$

with

$$|ED|e \le \lfloor 7\cdot 2^{n-2} - n-2\rfloor\ \omega^{2n}u|AD|e.$$

PROOF. Assume $n \ge 2$. Let $E = E^{(1)} + E^{(2)} +\ldots+ E^{(n-1)}$ where the $E^{(k)}$ are given by Lemma A.2. Substituting the bound on m_{ik} of Lemma B.1 into the bound on the elements $\varepsilon_{ij}^{(k)}$ we get

$$|\varepsilon_{ij}^{(k)}| \le \begin{cases} u|\bar{a}_{ij}^{(k)}| + 2\omega u|\bar{a}_{ik}^{(k)}| & \text{for } i,j > k, \\ u|\bar{a}_{ik}^{(k)}| & \text{for } i > j = k, \\ 0 & \text{otherwise.} \end{cases}$$

This implies

$$\sum_j |\bar{\varepsilon}^{(k)}_{ij}| \le (2n - 2k+1)\omega u|\bar{a}^{(k)}_{ik}| + u \sum_{j=k+1}^{n} |\bar{a}^{(k)}_{ij}|, \qquad i > k,$$

and from Lemma B.1 it follows that

$$\begin{aligned}\sum_j |\bar{\varepsilon}^{(k)}_{ij}| &\le (2n - 2k+1)\omega^{2k}u\left\{\sum_{\ell=1}^{k-1} 2^{k-1-\ell}|a_{i\ell}| + |\bar{a}_{ik}|\right\} \\ &\quad + (n-k)\omega^{2k-1}u \sum_{\ell=1}^{k-1} 2^{k-1-\ell}|\bar{a}_{i\ell}| + \omega^{k-1}u \sum_{j=k+1}^{n} |\bar{a}_{ij}| \\ &\le \begin{cases}(2n-1)\omega^2 u \sum_j |a_{ij}| & \text{if } k = 1, \\ (3n - 3k+1)\omega^{2k} u\, 2^{k-2} \sum_j |a_{ij}| & \text{if } k \ge 2.\end{cases}\end{aligned}$$

The lemma follows because

$$(2n-1) + \sum_{k=2}^{n-1} (3n - 3k+1)2^{k-2} = 7\cdot 2^{n-2} - n-2. \quad \text{Q.E.D.}$$

LEMMA B.3. *We have*

$$|\ell_{ij}\, \bar{u}_{ij}| \le \omega|\bar{a}^{(j)}_{ij}|,$$

$$|\bar{a}^{(j)}_{ij}| \le \omega^{2j-1} \sum_{\ell=1}^{j} \lceil 2^{j-\ell-1}\rceil |\bar{a}_{i\ell}|, \qquad i > j,$$

and

$$|\bar{u}_{ij}| \le |\bar{u}_{ii}|, \qquad i \le j.$$

PROOF. The first two inequalities follow immediately from Lemma B.1, and the third inequality is a consequence of column pivoting. Q.E.D.

LEMMA B.4. *The matrix* δU *of Lemma* A.5 *satisfies*

$$|L\, \delta U\, D|e \le (2^{n+1} - n-2)\omega^{2n}\, u|AD|e.$$

PROOF. Applying the inequality $|\bar{u}_{ij}| \le |\bar{u}_{ii}|$ of Lemma B.3 and the bounds of Lemma A.5, we get for $i < n$

$$\sum_j |\delta\bar{u}_{ij}| \le u \sum_{j=i}^{n} g(i,j)\omega^{n-i}|\bar{u}_{ii}|$$

$$\le \frac{(n-i+2)(n-i+1)}{2}\,\omega^{n-i}u|\bar{u}_{ii}|.$$

Therefore

$$(|L\delta UD|e)_i = \sum_j |\ell_{ij}| \sum_k |\delta\bar{u}_{jk}|$$

$$\le u \sum_j \frac{(n-j+2)(n-j+1)}{2}\,\omega^{n-j}|\ell_{ij}\;\bar{u}_{jj}|,$$

and so using Lemma B.3 gives

$$(|L\delta UD|e)_i \le \omega^{2n}u \sum_{j=1}^{n} \frac{(n-j+2)(n-j+1)}{2} \sum_{\ell=1}^{j} \lceil 2^{j-1-\ell}\rceil |\bar{a}_{i\ell}|$$

$$\le \omega^{2n}u \sum_{j=1}^{n} \frac{(n-j+2)(n-j+1)}{2} \lceil 2^{j-2}\rceil \sum_{\ell=1}^{n} |\bar{a}_{i\ell}|,$$

from which the lemma follows. Q.E.D.

LEMMA B.5. *The matrices* δU *and* δL *of Lemmas* A.5 *and* A.7 *satisfy* $|\delta L(U + \delta U)D|e \le 3(2^n-n-1)\omega^{2n}u|AD|e$.

PROOF. Applying the inequality $|\bar{u}_{ij}| \le |\bar{u}_{ii}|$ of Lemma B.3 and the bounds of Lemma A.5, we get

$$\sum_j |\bar{u}_{ij} + \delta\bar{u}_{ij}| \le \sum_{j=i}^{n} \omega^{n-i+1}|\bar{u}_{ii}|$$

$$\le (n-i+1)\omega^{n-i+1}|\bar{u}_{ii}|.$$

Therefore applying the bounds of Lemma A.7 gives

$$(|\delta L(U + \delta U)D|e)_i = \sum_j |\delta\ell_{ij}| \sum_k |\bar{u}_{jk} + \delta\bar{u}_{jk}|$$

$$\le \sum_j \min\{n-j+1,\ n-1\}\omega^{n-j}|a_{ij}^{(j)}/a_{jj}^{(j)}|(n-j+1)\omega^{n-j+1}|\bar{u}_{jj}|,$$

and so using Lemma B.3 gives

Appendix C. Error Bounds for Iterative Improvement

LEMMA C.1. *The computed residual satisfies*

$$r_m = b - Ax_m + c_m$$

with

$$|c_m| \leq u|b - Ax_m| + w|A||x_m|$$

where

$$w = \begin{cases} (1 + \dot{u})\ [(1 + \dot{u})^n - 1] & \textit{for s.p. accum.} \\ (1 + u)\ [(1 + u^2)^n - 1] & \textit{for d.p. accum.} \end{cases}$$

and $\dot{u} = u/(1 + u)$.

PROOF. The computed residual

$$r_m = b - (Ax_m + c'_m) + c''_m$$

where c'_m is the error due to computing Ax_m and c''_m is the rest of the error. For the single precision case we have

$$|c'_m| \leq [(1 + \dot{u})^n - 1]\ |A||x_m|$$

and

$$|c''_m| \leq \dot{u}|b - Ax_m - c'_m|$$

so that

$$\begin{aligned} |c_m| &\leq |c'_m| + |c''_m| \\ &\leq \dot{u}|b - Ax_m| + (1 + \dot{u})\ [(1 + \dot{u})^n - 1]\ |A||x_m|. \end{aligned}$$

For the double precision case we have

$$|c'_m| \leq [(1 + u^2)^n - 1]\ |A||x_m|$$

and

$$|c''_m| \leq u|b - Ax_m - c'_m|$$

so that

$$\begin{aligned} |c_m| &\leq |c'_m| + |c''_m| \\ &\leq u|b - Ax_m| + (1 + u)\ [(1 + u^2)^n - 1]\ |A||x_m|. \quad \text{Q.E.D.} \end{aligned}$$

$$(|\delta L(U + \delta U)D|e)_i \leq \omega^{2n} u \sum_{j=1}^{n} \min\{n-j+1, n-1\}(n-j+1) \sum_{\ell=1}^{j} \lceil 2^{j-1-\ell}\rceil |\bar{a}_{i\ell}|$$

$$\leq \omega^{2n} u \sum_{j=1}^{n} \min\{n-j+1, n-1\}(n-j+1) \lceil 2^{j-2}\rceil \sum_{\ell=1}^{n} |\bar{a}_{i\ell}|,$$

from which the theorem follows. Q.E.D.

THEOREM 5.1. *Let the vector* $\hat{x}$ *be computed by Gaussian elimination with column pivoting and column scaling where* $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$ *is the matrix of scale factors. Then there exists* ΔA *such that*

$$(A + \Delta A)\hat{x} = b$$

with

$$|\Delta AD|e \leq \tilde{\chi}(n)u|AD|e$$

where

$$\tilde{\chi}(n) = \lfloor 27 \cdot 2^{n-2} - 5n - 7\rfloor e^{2nu}.$$

PROOF. The theorem follows from the bounds of Lemmas B.2, B.4, and B.5 and from the equality

$$(A + E + \delta LU + (L + \delta L)\delta U)\hat{x} = b. \quad \text{Q.E.D.}$$

LEMMA C.5. *We have*

$$\overline{\lim_{m\to\infty}}\ \|z_m\| \leq \frac{uv + \tilde{w}}{1 - \tau} \|\,|A|\,|x|\,\|$$

provided that $\tau < 1$ *where*

$$\tau = u + (2v + uv + \tilde{w})\gamma.$$

PROOF. From Lemma C.4

$$\|z_{m+1}\| \leq \frac{\tau - v\gamma}{1 - v\gamma} \|z_m\| + \frac{uv + \tilde{w}}{1 - v\gamma} \|\,|A|\,|x|\,\|. \quad \text{Q.E.D.}$$

LEMMA C.6. *We have*

$$\overline{\lim_{m\to\infty}}\ |z_m| \leq \frac{\tilde{w}}{1 - u} \{|A|\,|x| - \|\,|A|\,|x|\,\|e\} + \frac{uv + \tilde{w}}{1 - \tau} \|\,|A|\,|x|\,\|e$$

provided that $\tau < 1$.

PROOF. From Lemma C.4

$$\overline{\lim_{m\to\infty}}\ |z_m| \leq u\ \overline{\lim_{m\to\infty}}\ |z_m| + \tilde{w}\{|A|\,|x| - \|\,|A|\,|x|\,\|e\}$$

$$+ \{(\frac{\tau - v\gamma}{1 - v\gamma} - u)\ \overline{\lim_{m\to\infty}}\ \|z_m\| + \frac{uv + \tilde{w}}{1 - v\gamma} \|\,|A|\,|x|\,\|\}e,$$

and the lemma follows from Lemma C.5. Q.E.D.

THEOREM C.7. *Assume* $|A|\,|x| > 0$. *Then*

$$\overline{\lim_{m\to\infty}}\ \eta_m \leq \frac{\frac{(1 + u\gamma)(uv + \tilde{w})\sigma}{1 - \tau} - \frac{\tilde{w}(\sigma - 1)}{1 - u} + \dot{u}}{1 - u - \frac{(1 + u)(uv + \tilde{w})\gamma\sigma}{1 - \tau}}$$

provided that $(1 + u)(uv + \tilde{w})\gamma\sigma < (1 - u)(1 - \tau)$ *where* $\sigma = \sigma_R(A, x)$.

PROOF. From Lemmas C.5 and C.6 we have

$$\overline{\lim_{m\to\infty}}\ \|z_m\|e \leq \frac{(uv + \tilde{w})\sigma}{1 - \tau} |A|\,|x|$$

and

$$\overline{\lim_{m\to\infty}}\ |z_m| \leq \{\frac{(uv + \tilde{w})\sigma}{1 - \tau} - \frac{\tilde{w}(\sigma - 1)}{1 - u}\}|A|\,|x|.$$

Hence using Lemma C.3,

LEMMA C.2. *Define* $z_m = A(x_m + d_m - x)$. *Then*

$$|z_m| \leq u|A(x_m - x)| + w|A||x_m|$$
$$+ (1 - v\gamma)^{-1}v\ (\| |A||x_m - x| \| + \gamma u\| |A(x_m - x)| \| + \gamma w\| |A||x_m| \|)e$$

where $v = \chi(n)u$ *and* $\gamma = \mathrm{Cond}(A^{-1})$.

PROOF. The correction term $d_m = (A + F_m)^{-1}r_m$ where F_m is the ΔA of Theorem 4.1. We have

$$z_m = A(x_m - x) + r_m - (I + F_m A^{-1})^{-1}\ F_m A^{-1} r_m$$
$$= c_m - (I + F_m A^{-1})^{-1}\ F_m(x_m - x - A^{-1}\ c_m).$$

It follows from the bounds of Theorem 4.1 that

$$z_m \leq |c_m| + (1 - v\gamma)^{-1}v\ (\| |A||x_m - x| \| + \gamma\|c_m\|)e.$$

Substituting the bounds of Lemma C.1 into this proves the lemma. Q.E.D.

LEMMA C.3. *We have*

$$|A(x_{m+1} - x)| \leq |z_m| + \dot{u}|A||x| + u\gamma\|z_m\|e$$

and

$$|A||x_{m+1} - x| \leq u|A||x| + (1 + u)\gamma\|z_m\|e.$$

PROOF. The new iterate

$$x_{m+1} = x_m + d_m + g_m$$

where $|g_m| \leq \dot{u}|x + d_m|$. Equivalently

$$x_{m+1} = x + A^{-1}z_m + g_m$$

where $|g_m| \leq \dot{u}|x| + \mathring{u}|A^{-1}z_m|$, from which the lemma follows. Q.E.D.

LEMMA C.4. *We have*

$$|z_{m+1}| \leq u(|z_m| - \|z_m\|e) + \tilde{w}(|A||x| - \| |A||x| \|e)$$
$$+ (1 - v\gamma)^{-1}\ \{[u + \gamma(v + uv + \tilde{w})]\|z_m\| + (uv + \tilde{w})\| |A||x| \|\}e$$

where $\tilde{w} = u^2 + w(1 + u)$.

PROOF. Substituting the bounds of Lemma C.3 into those of Lemma C.2 proves the lemma. Q.E.D.

PROOF. Substituting the inequality of Lemma C.8 into those of Lemma C.3 gives

$$|A(x_2 - x)| \leq (w + \dot{u})|A||x| + \{\frac{(v + u)w\gamma}{1 - v\gamma} + \frac{v(u + v\gamma + w)(1 + u\gamma)}{(1 - v\gamma)^2}\} \| |A||x| \|_e$$

and

$$|A||x_2 - x| \leq u|A||x| + \frac{(w + uv + v^2\gamma)}{(1 - v\gamma)^2} (1 + u)\gamma \| |A||x| \|_e,$$

and the theorem follows from the inequality

$$\eta_2 \leq \max \frac{|A(x_2 - x|}{|A||x| - |A||x_2 - x|} \quad . \quad \text{Q.E.D.}$$

$$\overline{\lim_{m\to\infty}} \, |A(x_m - x)| \le \{\frac{(1 + u\gamma)(uv + \tilde{w}\tau)\sigma}{1 - \tau} - \frac{\tilde{w}(\sigma - 1)}{1 - u} + \dot{u}\} \, |A|\,|x|$$

and

$$\overline{\lim_{m\to\infty}} \, |A|\,|x_m - x| \le \{u + \frac{(1 + u)(uv + \tilde{w})\gamma\sigma}{1 - \tau}\} \, |A|\,|x|.$$

The theorem follows from

$$\eta_m = \max \frac{|A(x_m - x)|}{|A|\,|x_m|} \le \max \frac{|A(x_m - x)|}{|A|\,|x| - |A|\,|x_m - x|} . \quad \text{Q.E.D.}$$

LEMMA C.8. *We have*

$$|z_1| \le w|A|\,|x| + \{\frac{wv\gamma}{1 - v\gamma} + \frac{v(u + v\gamma + w\gamma)}{(1 - v\gamma)^2}\} \, \|\,|A|\,|x|\,\| e.$$

PROOF. The first iterate $x_1 = (A + \Delta A)^{-1}Ax$,

and thus

$$x_1 - x = -A^{-1}(I + \Delta A\, A^{-1})^{-1}\Delta Ax.$$

So

$$\|A(x_1 - x)\| \le (1 - v\gamma)^{-1}v \, \|\,|A|\,|x|\,\|$$

and

$$\|\,|A|\,|x_1 - x|\,\| \le (1 - v\gamma)^{-1}v\gamma \, \|\,|A|\,|x|\,\|.$$

From Lemma C.2 we have

$$|z_1| \le w|A|\,|x|$$
$$+ (1 - v\gamma)^{-1} \{(v + w)\|\,|A|\,|x_1 - x|\,\| + u\|A(x_1 - x)\| + vw\gamma\|\,|A|\,|x|\,\|\}.$$

The lemma follows by substituting the previous two inequalities into this. Q.E.D.

THEOREM C.9. *Assume that* $|A|\,|x| > 0$. *Then*

$$\eta_2 \le \frac{w + \dot{u} + \frac{(v + u)w\gamma\sigma}{1 - v\gamma} + \frac{(u + v\gamma + w\gamma)(1 + u\gamma)v\sigma}{(1 - v\gamma)^2}}{1 - u - \frac{(w + uv + v^2\gamma)}{(1 - v\gamma)^2}(1 + u)\gamma\sigma}$$

provided that the denominator is positive.

BIBLIOGRAPHY

BAUER, F. L. Genauigkeitsfragen bei der Lösung linearer Gleichungssysteme. *ZAMM 46*, 7 (November 1966), 409-421.

BAUER, F. L. Computational graphs and rounding error. *SIAM J. Numer. Anal. 11*, 1 (March 1974), 87-96.

FORSYTHE, G. E., AND MOLER, C. B. *Computer Solution of Linear Algebraic Systems*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1967.

GEAR, C. W. Numerical errors in sparse linear equations. File F-75-885, Dept. of Comput. Sci., Univ. of Illinois at Urbana-Champaign, Urbana, Illinois, May 1975.

HAMMING, R. W. *Introduction to Applied Numerical Analysis*. McGraw-Hill, New York, 1971.

KAHAN, W. Numerical linear algebra. *Canad. Math. Bull. 9* (1966), 757-801.

MILLER, W. On the stability of finite numerical procedures. *Numer. Math 19* (1972), 425-432.

MILLER, W. Computer search for numerical instability. *J.ACM 22*, 4 (October 1975), 512-521.

MILLER, W. Roundoff analysis by direct comparison of two algorithms. *SIAM J. Numer. Anal. 13*, 3 (June 1976), 382-392.

SHERMAN, A. H. Algorithms for sparse Gaussian elimination with partial pivoting. Rpt. R-76-817, Dept. of Comput. Sci., Univ. of Illinois at Urbana-Champaign, Urbana, Illinois, July 1976.

STEWART, G. W. *Introduction to Matrix Computations*. Academic Press, New York, 1973.

VAN DER SLUIS, A. Stability of solutions of linear algebraic systems. *Numer. Math. 14* (1970), 246-251.

VAN DER SLUIS, A. Condition, equilibration, and pivoting in linear algebraic systems. *Numer. Math. 15* (1970), 74-86.

WILKINSON, J. H. *Rounding Errors in Algebraic Processes*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1963.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

19 REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
18 AFOSR-TR-77-~~0~~8~~0~~1

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER
9

4. TITLE (and Subtitle)
6 GAUSSIAN ELIMINATION AND NUMERICAL INSTABILITY,

5. TYPE OF REPORT & PERIOD COVERED
Interim rept.,

6. PERFORMING ORG. REPORT NUMBER
14 UIUCDCS-R-77-862, UILU-ENG-77-1714

7. AUTHOR(s)
10 ROBERT D. SKEEL

8. CONTRACT OR GRANT NUMBER(s)
15 AF-AFOSR-~~75~~-2854-75

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Computer Science
University of Illinois
Urbana, Illinois 61801

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
61102F 16 2304 A3 17

11. CONTROLLING OFFICE NAME AND ADDRESS
Air Force Office of Scientific Research/NM
Bolling AFB DC 20332

12. REPORT DATE
11 April 1977

13. NUMBER OF PAGES
62

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
12 67p.

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

numerical stability, Gaussian elimination, iterative improvement, ill conditioning, scaling, equilibration, pivoting

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Roundoff error in the solution of linear algebraic systems is studied using a more realistic notion of what it means to perturb a problem, namely that each datum is subject to a relatively small change. The condition number is determined for this approach. A good computable error bound is given for the "backward error." The effect of scaling on the stability of Gaussian elimination is studied, and it is discovered that the proper way to scale a system is dependent on knowing the solution. Finally it is shown that Gaussian elimination can be stabilized by doing iterative improvement.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

176 011

BIBLIOGRAPHIC DATA SHEET

1. Report No.
UIUCDCS-R-77-862

2.

3. Recipient's Accession No.

4. Title and Subtitle

GAUSSIAN ELIMINATION AND NUMERICAL INSTABILITY

5. Report Date
April 1977

6.

7. Author(s)
Robert D. Skeel

8. Performing Organization Rept. No. 862

9. Performing Organization Name and Address

Department of Computer Science
University of Illinois
Urbana, Illinois 61801

10. Project/Task/Work Unit No.

11. Contract/Grant No.

AFOSR-75-2854

12. Sponsoring Organization Name and Address

Air Force Office of Scientific Research/NM
Bolling AFB DC 20332

13. Type of Report & Period Covered

interim

14.

15. Supplementary Notes

16. Abstracts

Roundoff error in the solution of linear algebraic systems is studied using a more realistic notion of what it means to perturb a problem, namely that each datum is subject to a relatively small change. The condition number is determined for this approach. A good computable error bound is given for the "backward error." The effect of scaling on the stability of Gaussian elimination is studied, and it is discovered that the proper way to scale a system is dependent on knowing the solution. Finally it is shown that Gaussian elimination can be stabilized by doing iterative improvement.

17. Key Words and Document Analysis. 17a. Descriptors

numerical stability, Gaussian elimination, iterative improvement,
ill conditioning, scaling, equilibration, pivoting

17b. Identifiers/Open-Ended Terms

17c. COSATI Field/Group

18. Availability Statement

unlimited

19. Security Class (This Report)
UNCLASSIFIED

20. Security Class (This Page
UNCLASSIFIED

21. No. of Pages
65

22. Price